Evolution of mammalian genome architecture through retrotransposition

By

REUBEN MACKENZIE BUCKLEY

THE UNIVERSITY
of ADELAIDE

Department of Genetics and Evolution

School of Biological Sciences

A thesis presented for the degree of DOCTOR OF PHILOSOPHY

AUGUST 2017

Abstract

Retrotransposons, mobile DNA elements that replicate via a copy and paste mechanism, are a major component of mammalian genome architecture. They account for at least one-third of the human genome and are major drivers of lineage-specific gain and loss of DNA. While there are many examples of how specific retrotransposons have impacted evolution, their interaction with large-scale genome architecture remains poorly characterised. Throughout my thesis I investigated two fundamental questions regarding genome evolution and retrotransposons. Firstly, how does genome architecture shape retrotransposon accumulation? Secondly, how does retrotransposon accumulation in turn impact on genome architecture?

The current model of retrotransposon accumulation largely relies on local sequence composition. However, this model fails to account for genome-wide chromatin structure, an important factor that regulates DNA accessibility to insertion machinery. By analysing retrotransposon accumulation at open chromatin sites I showed that genome structure strongly associates with retrotransposon accumulation patterns. In addition, by mapping retrotransposon accumulation patterns of non-human mammals back to human, I was able to observe large-scale positional conservation of lineage-specific retrotransposons. These findings suggest that through conservation of synteny, gene regulation and nuclear organisation, retrotransposon accumulation in mammalian genomes follows similar evolutionary trajectories.

Beneath the conserved structural framework of mammalian genomes there exists a high degree of lineage-specific turnover of DNA. Outside of whole genome duplication, retrotransposons are the largest contributing factor to genome growth. In contrast to this, accumulation of retrotransposons can also increase the probability of unequal crossing over causing DNA loss through large deletion events. Using multiple pairwise alignments I calculated regional levels of lineage-specific DNA gain and loss in the human and mouse genomes. I found that while lineage-specific DNA loss overlapped with open chromatin regions in both genomes, different sources for lineage-specific DNA gain drove divergence in genome architecture. These findings reveal the turbulent nature of lineage-specific evolution of large-scale genome architecture, ultimately questioning the evolutionary stability of structural chromosomal domains.

In addition to analysing large-scale genome architecture I performed two separate analyses on retrotransposons in the bovine genome. Due to the presence of BovB retrotransposons, the bovine retrotransposon landscape is clearly distinct from other placental mammals. For the first analysis, I identified bovine-specific retrotransposon associated gene co-expression networks. Following the genomic distribution of bovine retrotransposons, my results show that gene expression strongly associates with genome architecture. For the second analysis, I characterised retrotransposons surrounding tandem duplicate copies of the bovine *NK-lysin* gene. My results were consistent with retrotransposon accumulation causing genomic rearrangements via non-allelic homologous recombination.

Altogether, my thesis reveals hidden interactions between retrotransposon accumulation, and mammalian genome structure and function. By re-purposing publicly available datasets I have characterised various aspects of the complex co-evolutionary relationships between retrotransposons and the genomes in which they reside in.

Dedication and Acknowledgements

On the 17^{th} of June 2013, I stepped into the Braggs 2^{nd} floor write-up room, sat down at my desk and began to work on what ultimately became my PhD thesis. Four years, two months, one week and six days later, I have submitted my thesis and would like to thank all of the people that helped make it happen.

To begin I would like to thank my PhD supervisors Dave and Dan. Together your supervision provided me with enough freedom and guidance that I could follow through with my own ideas, no matter how unorthodox they were, and still not get lost in the process. Dave, despite your incredibly busy schedule, I am grateful that you were always able to put your students first. Answering their emails and working on their manuscripts while your flights are getting cancelled and diverted is by no means a simple task. Dan, one of the many things I appreciated was Tuesday afternoon coffee time. I and my fellow students learnt things in those conversations that could not be taught in the lab. I am aware that the past few years have carried their own personal difficulties and I am extremely appreciative of all your efforts.

I would also like to thank my fellow lab mates. You have all been a great group to work with and I have been glad to know you over the years. For special mention I'd like to acknowledge my longest serving fellow lab mate Zhipeng. You have always been able to offer me thoughtful advice and encouragement when I needed it most. As for my fellow bioinformatics PhD candidates Lu and Atma, I am not sure where to begin. For starters, I will forever be amazed by Lu's ability to deliver obscure and unique backhanded compliments. These have often left me speechless, dumbfounded or just plain confused. At the same time, Atma was able to stay out of the firing line and still get a front row seat to the surrounding conflict, sometimes even spurring it on. In all seriousness you guys made my candidature a fun and enjoyable experience. By the way, I haven't forgotten about that bet we made! For their help and support throughout 2016 and 2017 I would like to thank Brittany and James. Both of you were to work with as practical demonstrators and made excellent lab mates. I wish you guys all the best with your respective upcoming PhD candidature opportunities. In addition, special thanks goes out to the new recruits Catisha and Urwah, particularly for their invaluable assistance with editing my many thesis drafts. Finally, I would like to give an honourable mention to Armstrong (the unofficial lab member). I have had the pleasure of getting to know you over the past few years as someone I can speak with candidly. While we occasionally disagree, I feel as though our

many discussions have helped me develop my arguments and sharpen my rhetoric.

Importantly, I will take this opportunity to pass on my gratitude to my many friends who have helped me outside of the lab. Jonny and Ernesto, the other two thirds of what I consider the original honours/PhD trio. You have both expanded my horizons in what can be considered 'edible' in terms of birthday cake. The CBC guys: Anton, Ray, Billo, Josh, Tom, Stefan, Carmine and Imre, we've all known each other for almost fourteen years now. You guys have always been there for me whenever I needed a distraction from the harrows of PhD life. When it comes to relaxing, eating junk food and watching awesome movies, I can't look past Josy and Cam, you guys provided excellent company and a guilt-free environment. Finally, I'd like to acknowledge my previous house-mate Danna. You have helped me through some of my most difficult times and have always been there in one way or another. Knowing you throughout the years has helped me to both centre my thoughts and broaden my perspectives. I wish all of you the best in your future endeavours.

Last but by no means least I would like to thank my family. You had to deal with me at some of my lowest points. Your unwavering support throughout the last four years has meant, and will continue to mean, the world to me. I would especially like to thank my mother Renee. I know it has been a difficult year and that having your twenty-six year old son move back home to finish his thesis creates a less than ideal situation. Your ability to make the best out of any situation and simultaneously commit yourself to taking care of your many children is truly inspiring. I am grateful for how you and Dad raised me, and one day I hope to be as productive, thoughtful and considerate as you are.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....12/12/2017.....

Table of Contents

Chapter 1

Introduction

Genome evolution in complex organisms is by no means a straight forward process. For example, protein-coding genes make up less than 2% of the human genome and are highly conserved across mammals. In contrast, the remaining 'non-coding' fraction of the genome contains all the necessary information required for regulating complex systems of proteins and RNA molecules. However, the non-coding portion of the genome is highly dynamic, where estimates - along with various definitions - of how much DNA is actually 'functional' differ widely. At the centre of this conundrum are retrotransposons; self-replicating mobile segments of DNA. They have the potential to cause large mutations but can also act as gene regulatory elements. In this chapter I provide an overview of the mammalian retrotransposon landscape and review the literature regarding mammalian epigenetic retrotransposon silencing mechanisms. I discuss how these silencing mechanisms can be co-opted along with retrotransposons themselves to shape mammalian gene regulatory networks and impact evolution.

Statement of Authorship

Title of Paper	Mammalian genome evolution as a result of epigenetic regulation of transposable elements
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Buckley, Reuben M., and David L. Adelson. "Mammalian genome evolution as a result of epigenetic regulation of transposable elements." *Biomolecular concepts* 5.3 (2014): 183-194.

Principal Author

Name of Principal Author (Candidate)	Reuben Buckley		
Contribution to the Paper	Analysed literature, prepared figures and wrote manuscript.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	27/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of work and assisted in writing the manuscript		
Signature		Date	25/8/2017

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Review

Reuben M. Buckley and David L. Adelson*

Mammalian genome evolution as a result of epigenetic regulation of transposable elements

Abstract: Transposable elements (TEs) make up a large proportion of mammalian genomes and are a strong evolutionary force capable of rewiring regulatory networks and causing genome rearrangements. Additionally, there are many eukaryotic epigenetic defense mechanisms able to transcriptionally silence TEs. Furthermore, small RNA molecules that target TE DNA sequences often mediate these epigenetic defense mechanisms. As a result, epigenetic marks associated with TE silencing can be reestablished after epigenetic reprogramming – an event during the mammalian life cycle that results in widespread loss of parental epigenetic marks. Furthermore, targeted epigenetic marks associated with TE silencing may have an impact on nearby gene expression. Therefore, TEs may have driven species evolution *via* their ability to heritably alter the epigenetic regulation of gene expression in mammals.

Keywords: epigenetics; genome evolution; mammals; transposable elements.

DOI 10.1515/bmc-2014-0013
Received April 7, 2014; accepted May 27, 2014

List of abbreviations: AGO2, Argonaute 2; APOBEC3, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3; ChIP, chromatin immunoprecipitation; ChIP-seq, ChIP sequencing; CTCF, CCCTC binding factor; DNMT1, DNA methyltransferase 1; DNMT3a, DNA methyltransferase 3a; DNMT3b, DNA methyltransferase 3b; DNMT3L, DNA methyltransferase 3-like; dsRNA, double-stranded RNA; ERVs, endogenous retroviruses; ESC, embryonic stem cell; ESR1, estrogen receptor 1; G9a, euchromatic histone-lysine *N*-methyltransferase 2; H3K27ac, histone H3 lysine 27 acetylation; H3K27me3, histone H3 lysine 27 methylation 3; H3K4me1, histone H3 lysine 4 methylation 1; H3K4me3, histone H3 lysine 4 methylation 3; H3K9, histone H3 lysine 9; H3K9me2/3, histone H3 lysine 9 methylation 2/3; IAP, intracisternal A particle; kb, kilobases; LINEs, long interspersed nuclear elements; LTR, long terminal repeat; MILI, piwi-like RNA-mediated gene silencing 2; miRNA, micro-RNA; MIWI2, piwi-like RNA-mediated gene silencing 4; MOV10, Moloney leukemia virus 10, homologue; MOV10L1, MOV10-like 1; ORFs, open reading frames; PGCs, primordial germ cells; piRNAs, PIWI-interacting RNAs; Pld6, phospholipase D family, member 6; priRNAs, primary RNAs; RdDM, RNA-directed DNA methylation; RNAi, RNA interference; RNP, ribonucleoprotein; SETDB1, SET domain bifurcated 1; SINEs, short interspersed nuclear elements; siRNA, small interfering RNA; sRNA, small RNA; Suv39, suppressor of variegation 3–9; TEs, transposable elements; TFBSs, transcription factor binding sites; TFs, transcription factors; UHRF1, ubiquitin-like with PHD and ring finger domains 1.

Introduction

Transposable elements (TEs) are mobile DNA segments that have had an extensive effect on mammalian genome evolution (1). As much as two thirds of the human genome may be composed of repetitive sequences, of which TE-derived sequences are a major component (2). Because of their ability to replicate themselves and their potential to cause mutation *via* insertional mutagenesis or ectopic recombination resulting in large genomic rearrangements, TEs have long been thought of as selfish genetic elements (3–5). This view is also consistent with the significant role TEs have been shown to play in various diseases (6, 7). Genome defense from TEs is largely mediated by transcriptional silencing. This is achieved by epigenetic modifications that disrupt the accessibility of the necessary transcriptional machinery. Both DNA methylation and

***Corresponding author: David L. Adelson,** School of Molecular and Biomedical Science, University of Adelaide, North Terrace, Adelaide, South Australia 5005, Australia,
e-mail: david.adelson@adelaide.edu.au
Reuben M. Buckley: School of Molecular and Biomedical Science, University of Adelaide, North Terrace, Adelaide, South Australia 5005, Australia

histone modifications are involved in these processes and are mediated by RNA intermediates (8–11).

However, evidence is emerging that suggest TEs are more than just genomic parasites. TE insertions have been shown to affect nearby gene expression in a variety of ways. Examples include TEs providing alternative splice sites, transcription factor binding sites (TFBSs), and alternative promoters for genes [reviewed in (1)]. Interestingly, epigenetic silencing mechanisms associated with TEs also affect gene expression. A well-studied example of this phenomenon is epigenetic inheritance at the *axin-fused* allele in which a kinky tail phenotype associates with differential methylation of the long terminal repeat (LTR) at the 3′ end of an intracisternal A particle (IAP) element in mice. Hypermethylation of the 3′ LTR of the IAP element was shown to suppress the kinked tail phenotype by silencing a cryptic promoter. Crosses between penetrant and silent *axin-fused* mice with null mice showed that the penetrance of the allele was inherited. This implied that the epigenetic methylation state of the IAP element remained stable as it passed from one generation to the next. Furthermore, the epigenetic state of an individual's sperm cells reflected the epigenetic state of that individual's somatic cells, thereby providing a mechanism for inheritance (12).

In the above example, the epigenetic state associated with the IAP element was inherited by the next generation and also affected expression of the *axin-fused* allele. Transgenerational inheritance of stable patterns, such as DNA sequences, provides the foundations on which evolutionary processes such as natural selection act. Although IAP elements appear to be an exception to most TEs, due to their ability to avoid epigenetic reprograming, the RNA intermediates that target TEs for transcriptional silencing through epigenetic modifications provide a mechanism by which epigenetic patterns associated with TEs can be inherited (13, 14). In this context, TEs can be thought of as providing a unique epigenetic environment. Throughout this review, we explore the role TEs have played in altering the epigenetic landscape, which in turn, may have altered gene expression patterns and regulatory networks and thereby driven evolution in different species.

The mammalian TE landscape

To understand the potential evolutionary impact of TEs, we must take into account the various types and families of TEs with different ages, mechanisms of action, distributions across species, and distributions within genomes. In this section, we discuss how each of these factors shapes the mammalian TE landscape.

According to the Repbase classification system, there are two main types of TEs: type 1 and type 2. Type 1 TEs consist of LTR and non-LTR retrotransposons. Many LTR retrotransposons in mammals are endogenous retroviruses (ERVs). ERVs have been grouped into several different classes based on such criteria as structural features and phylogeny [reviewed in (15)]. Non-LTR retrotransposons are made up of long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (16). LINEs are usually several kilobases long and contain two open reading frames (ORFs), one of which encodes a ribonucleoprotein (RNP) that reverse transcribes the element and inserts the DNA copy into the genome. The copy and paste process of L1 retrotransposition is shown in Figure 1. However, SINEs are only approximately 300 base pairs (bp) long, contain no ORFs, and require the retrotransposition machinery encoded by LINE elements for retrotransposition. SINEs are derived from the 3′ end of LINEs, and these 3′ sequences bind the LINE-encoded RNP required for replication. LINEs and their derived SINEs are referred to as LINE-SINE pairs. In humans, LINE L1 and SINE Alu are an active LINE-SINE pair. In the mouse, a similar pairing also exists, where mouse LINE L1s form a LINE-SINE pair with SINE B1 elements (17). However, the majority of TE sequences in mammalian genomes are inactive. Type 2 TEs, also known as DNA transposons, are able to excise themselves from the genome and reinsert themselves elsewhere in the genome using a transposase encoded in their single ORF. Because of this cut-and-paste mobilization that does not generate additional copies, type 2 TEs are found in much lower numbers than type 1

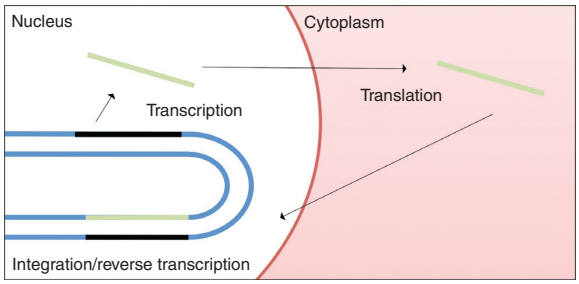


Figure 1 Retrotransposition.
Class 1 non-LTR TEs (black) are transcribed (green) and transported to the cytoplasm. Within the cytoplasm, non-autonomous TEs undergo translation and produce an RNP. TE transcripts are transported back into the nucleus where they are reverse transcribed and integrated into the genome. The above process has resulted in large portions of the genome comprising of repeated DNA sequences.

TEs in mammalian genomes. The above findings have previously been reviewed by Jurka et al. (18).

The mammalian TE landscape is very complex, and every species of mammal has both shared and unique TEs that can be traced back to various lineages within the mammalian radiation. Initial genome-wide studies of TE distribution based on the human genome concluded that LINE L1s were more prevalent in AT-rich regions; SINE MIRs and SINE Alus showed a preference for GC-rich regions; and LINE L2s were distributed independent of GC content (19, 20). LINE L1s, LINE L2s, and SINE Alus also all had a preference for antisense insertions within genes; this was most pronounced for LINE L1s. SINE MIR sequences, however, showed no such insertion preference. This observation was interpreted as the result of selection against LINE L1 sense insertions because the LINE L1 element's poly A signal/tail may cause shortened gene transcripts (19). Therefore, the observed TE distribution is the product of both TE insertion preference and selection against specific types of TE insertions (19). In the mouse, the TE landscape is very different. For example, young SINE B1 and SINE B2 elements insert into SINE-rich GC areas, whereas young SINE Alus in human insert into SINE Alu-poor AT-rich areas. Human and mouse also differ in retrotransposon content. For example, the human genome has fewer LTR/ERVs compared with the mouse genome, and the mouse genome has far fewer SINE MIRs and LINE L2s than the human genome (21).

Throughout the mammalian lineage, older TEs show signs of being retained, which result from selective pressures. For example, SINE MIRs and LINE L2s and TE-free regions are often found in conserved orthologous segments between human and mouse (22, 23). Moreover, subsequent analyses of repeat families in different species have adopted a more global approach to identify associations of repetitive elements in different families across species. This led to the identification of regions enriched for ancestral repeats (SINE MIR and LINE L2) in human, horse, and cow. Therefore, ancestral mammalian TEs show signs of both positional and sequence conservation in a number of species (24, 25). This conservation suggests a role for these repeats in the genome structure associated with the regulation of gene expression.

Although distantly related species have been used to compare the distribution of inactive and ancestral repeats, comparisons between closely related species have been used to compare distributions of young, active TEs. Deep sequencing of 17 strains of mouse revealed over 100,000 TE variants, each of which had survived selection over the past 2 million years. The ERV family of repeats underwent the largest expansion, and deleterious ERVs were rapidly purged from the mouse genome. Deleterious LINE L1s were also purged but not quite as rapidly as ERVs. ERV insertions were also shown to be most highly associated with changes in gene expression between the mouse strains (26). It is clear that TEs are a source of variation among species and can cause large genomic changes. However, most such changes are detrimental, and it is therefore advantageous to be able to reduce the probability of potentially detrimental changes.

Silencing of TEs *via* targeted epigenetic mechanisms

TE silencing through DNA methylation and chromatin modification keeps retrotransposition in check by suppressing TE transcription. However, during germ cell early embryonic development, DNA and histone methylation patterns are transiently erased, allowing TEs to mobilize. However, mobilizing TEs are quickly inhibited by targeted small RNA (sRNA) TE-silencing mechanisms (27, 28). After this transient demethylation, DNA methyltransferase 3a (DNMT3a) and DNA methyltransferase 3-like (DNMT3L) specifically methylate TEs, thereby suppressing TE transcription (8, 29, 30).

The most well-characterized sRNA-targeting mechanism for TE transcriptional silencing in mammals is the PIWI-based recognition system (27). sRNA molecules approximately 26–31 nucleotides long direct DNA methylation at TE promoters in a general process known as RNA-directed DNA methylation (RdDM), and these RNA molecules are known as PIWI-interacting RNAs (piRNAs) (11). Primary piRNAs are generated from piRNA clusters during widespread TE transcription during epigenetic reprogramming. Primary piRNAs then bind to piwi-like RNA-mediated gene silencing 2 (MILI) to form complexes that then bind to and cleave the TE transcripts. The cleaved transcript product is a secondary piRNA that forms a complex with piwi-like RNA-mediated gene silencing 4 (MIWI2) and targets the primary piRNA cluster transcript, thereby leading to increased production of primary piRNAs. This process is known as a 'ping-pong' amplification cycle and is very effective in dealing with large numbers of TE transcripts (Figure 2). The PIWI-based recognition system occurs before DNMT3L-guided methylation and is believed to be the causal factor in TE methylation specificity (11). RNA intermediates may also be involved in directing chromatin modifications that can silence TE transcription. This idea is well established in plants and supported in *Drosophila* but has not

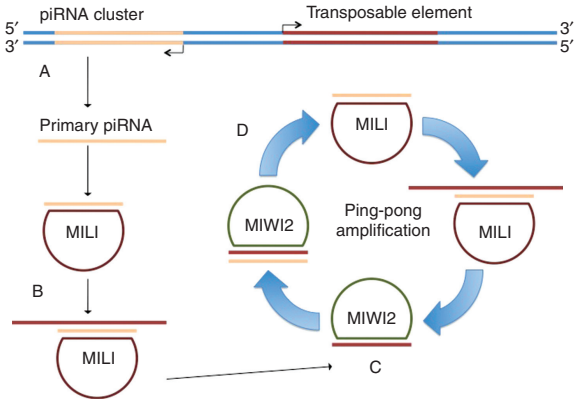


Figure 2 piRNAs and ping-pong amplification.
(A) piRNAs targeting TEs are transcribed from piRNA clusters. (B) Primary piRNAs are processed and loaded into MILI where they are able to guide MILI to TE transcripts, initiating the formation of the ping-pong amplification pathway. (C) Cleaved TE transcripts become secondary piRNAs and are loaded into MIWI2. (D) The complex then targets and cleaves primary piRNA clusters, resulting in the generation of more piRNAs loaded into MILI. piRNAs generated by this process are believed to drive RdDM by an unknown mechanism.

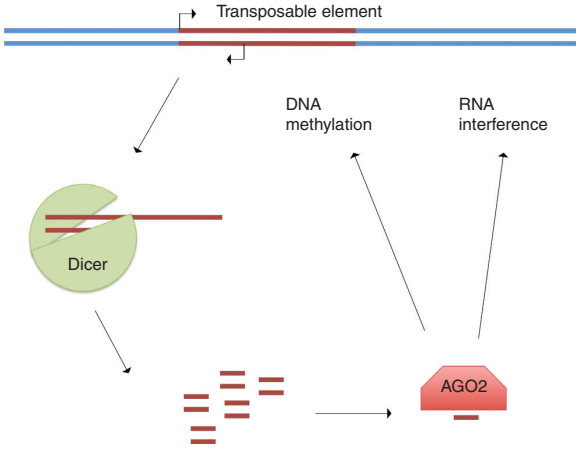


Figure 3 TE silencing *via* sRNA molecules.
Double-stranded TE RNA is produced as a result of transcription from bidirectional TE promoters. This dsRNA is then targeted and cleaved by DICER. The resulting sRNAs are then loaded into AGO2 and direct DNA methylation or RNA interference.

been confirmed in mammals (31, 32). piRNA sequences are found in clusters throughout the genome and share sequence similarity with TEs. The production of piRNAs corresponding to a specific TE is likely the result of a TE insertion into a piRNA cluster (33, 34). Although piRNAs play a large role in silencing TEs, some TE families are effectively silenced even in their absence. For example, SINE B1 elements in mouse have a locus-to-locus variation in their methylation patterns. Knockout of phospholipase D family, member 6 (*Pld6*), and *MILI* genes, which are involved in piRNA biogenesis, results in disrupted piRNA-mediated silencing of LINE L1 elements, whereas methylated SINE B1 elements remain methylated in spermatogonia (35, 36). The knockouts also show no increase in SINE B1 expression, indicating that SINE B1 silencing occurs independently of piRNA activity (36).

Recent work on LINE L1 silencing shows the potential involvement of other RdDM mechanisms in mammals. Mammalian micro-RNAs (miRNAs) associated with repeats are 22 nucleotides long and are products of double-stranded RNA (dsRNA) cleaved by DICER and loaded into Argonaute 2 (AGO2) (37, 38) (Figure 3). Mouse embryonic stem cell (ESC) *DICER* knockouts showed that mammalian repeat-associated miRNAs were depleted, LINE L1 promoter elements were hypomethylated, and that LINE L1 transcription, translation, and copy number had increased. Therefore, components of miRNA biogenesis in mammals are linked to transcriptional silencing of TEs (39, 40). Furthermore, sRNAs in mammals are involved with RNA interference (RNAi) or post-transcriptional silencing of TEs. dsRNA processed by DICER caused sRNAs in mammals to form perfect small interfering RNA (siRNA) duplexes with two-nucleotide 3′ overhangs, a characteristic associated with RNAi in other systems (40).

Post-transcriptional processing of TE transcripts involves several other regulators. The microprocessor, a multiprotein complex able to recognize and cleave primary RNAs (priRNAs), plays an important role in miRNA biogenesis (41, 42). The microprocessor recognizes structures within LINE L1 elements and promotes their degradation (43). Another regulator of TE activity through RNAi is the human RNA helicase, Moloney leukemia virus 10, homologue (MOV10) (44, 45). MOV10 is part of a multiprotein complex with other components involved in RNA-induced silencing (46). MOV10 prevents retrotransposition of non-autonomous TEs by interacting with the LINE L1 RNP (45). Furthermore, a protein related to MOV10 known as MOV10-like 1 (MOV10L1), interacts with piRNA proteins in male mouse germ cells and is involved in transcriptional silencing of certain TE families (47, 48). MOV10 is also found in complexes with members of the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) family (APOBEC3G and APOBEC3F) and is associated with defense against retroviruses, which have replication mechanisms similar to retrotransposons (49, 50). Therefore, APOBEC3 proteins may also be involved in various processes that protect host genomes against TEs (51). The APOBEC3 family of proteins is a family of

cytidine deaminases that convert cytidine to uridine to edit retrotransposon DNA and cDNA as a defense against retrotransposition (51).

Histone modifications provide another mechanism to regulate TE expression. Most active TE sequences are associated with histone H3 lysine 9 (H3K9) methylation and are therefore transcriptionally repressed (52). For example, mutations in methyltransferases that are associated with repressive histone modifications lead to increased TE activity (53). In mouse early embryogenesis, a methyltransferase known as SET domain bifurcated 1 (SETDB1) targets specific promoter-proximal class I and class II ERVs. Embryonic cells lacking SETDB1 show transcription of the promoter-proximal ERVs in the form of aberrant gene transcripts that would otherwise be silenced. Therefore, SETDB1 is involved with transcriptional control of TEs independent of DNA methylation (10). Another methyltransferase found in mouse involved in TE silencing is euchromatic histone-lysine *N*-methyltransferase 2 (G9a) (54). G9a does not appear to be involved with silencing class I and II ERVs but is necessary for silencing class III ERVs (55). Suppressor of variegation 3–9 (Suv39) is another H3K9 methyltransferase also linked to TE silencing; deletions of *Suv39* result in a moderate increase in TE activity (52). Furthermore, the above mechanisms may hold for a variety of mammals including the pig. For example, porcine ERVs are silenced by similar chromatin modifications as seen in mouse (56). However, heterochromatic silencing during embryogenesis may not be an active driver of TE silencing. Moreover, TE silencing usually occurs after loss of an active histone mark and before gain of a repressive histone mark (57).

Interestingly, heterochromatin modifications associated with TE sequences may be selected to play a dual role, resulting in further downstream implications of TE accumulation. Generally, heterochromatin causes a transcriptionally repressive environment-reducing TE activity. However, heterochromatic regions are also unable to undergo recombination (3). This is important because unregulated TE genomic-enriched regions are particularly prone to hazardous recombination events that have been linked to disease in humans (5, 7). Therefore, prevention of non-homologous recombination may be a driving force behind heterochromatic repression of TEs. In addition, recombination also often results in TE deletion, implying that heterochromatic silencing may be the cause of TE accumulation (4). Furthermore, simulations have shown that under an ectopic recombination model, TEs accumulate in regions of low recombination (58).

This section shows how epigenetic mechanisms are involved in the regulation of TEs. Epigenetics are now known to contribute to many regulatory processes, especially throughout development. The following section aims to show the breadth of developmental regulation under the control of epigenetic processes in the context of the regulatory impact of retrotransposition.

Epigenetic regulatory mechanisms are essential for mammalian development

Epigenetic mechanisms are well characterized in terms of the roles they play in development. Mammalian development is highly complex and requires extensive regulation of intricate cellular processes. During development, the differentiation potential of cells is gradually reduced at each stage until cells terminally differentiate. This reduction of differentiation potential is largely regulated by epigenetics.

DNA methylation is the critical modification of DNA with respect to the epigenetic regulation of transcription. Specifically, DNA methylation refers to the methylation of cytosine and occurs mostly in CpG sequences (59). In mammals, approximately 60%–80% of CpGs are methylated. However, approximately 10% of CpGs are resistant to methylation and are found in GC-rich regions of the genome. These CpG sites are known as CpG islands and are found in gene and retrotransposon promoters (28, 60). The DNA methyltransferases, DNA methyltransferase 1 (DNMT1), DNMT3a, and DNA methyltransferase 3b (DNMT3b) all have roles in maintaining DNA methylation throughout mammalian development (61, 62). The deletion of *DNMT1* in ESCs results in apoptosis, whereas simultaneous deletion of *DNMT3a* and *DNMT3b* did not affect survivability yet resulted in ESCs unable to differentiate (63). Once established, methylation patterns are able to persist through multiple rounds of mitosis. During the S phase, DNMT1 directly interacts with proliferating cell nuclear antigen and ubiquitin-like with PHD and ring finger domains 1 (UHRF1); this complex is recruited to sites of DNA replication and binds hemi-methylated DNA *via* a SET- and RING-associated domain (64–66). UHRF1 binds to parental methylated DNA and thereby directs DNMT1 to the daughter strand (28, 66). Therefore, DNA methylation is a stable process for transmitting epigenetic regulatory information from parent to daughter cell, unlike transmission of epigenetic regulatory information from parent to offspring at the level of multicellular organisms.

Inherited information from a parent to offspring is largely mediated by one cell, a single gamete. Gametes from each parent fuse to form a zygote, which then develops into an organism made up of a large variety of tissues and differentiated cell types. For this process to occur properly, there are two stages during development where cells undergo epigenetic reprogramming resulting in global hypomethylation. The processes governing how epigenetic patterns are reestablished during development are complex and remain an area of intense research. In ESCs and primordial germ cells (PGCs), epigenetic states are reset requiring that methylation patterns are reestablished in a targeted manner for differentiation to occur. Various DNA methylation target sites have been identified. These include promoters, pericentromic repeats, TEs, and imprint control regions (28).

Another form of epigenetic regulation during mammalian development is through histone modifications. Histone proteins form a complex with DNA called a nucleosome, in which approximately 147 nucleotides of DNA are wrapped around the nucleosomal histones H2A, H2B, H3, and H4. Two copies of each histone make up the nucleosome, and a collection of nucleosomes results in the formation of chromatin. Each one of these histones can also be chemically modified, usually by a methyltransferase or an acetylase. Chemical modifications of histone proteins regulate the accessibility of surrounding DNA (67). For example, repressive histone modifications cause nucleosomes to tightly associate, resulting in the surrounding DNA being made inaccessible and transcriptionally silent. Known repressive histone modifications include histone H3 lysine 9 methylation 2/3 (H3K9me2/3) and histone H3 lysine 27 methylation 3 (H3K27me3) (68–70). Meanwhile, active histone modifications can cause the nucleosomes to dissociate, resulting in the surrounding DNA becoming accessible to transcriptional machinery. Active histone modifications at promoters include histone H3 lysine 4 methylation 3 (H3K4me3), histone H3 lysine 27 acetylation (H3K27ac), and known modifications at enhancers include H3K27ac and histone H3 lysine 4 methylation 1 (H3K4me1) (70–72). These histone modifications cause a change in chromatin status at particular loci; however, they are not as stable as DNA methylation (67). Therefore, the DNA loci associated with histone modifications that are analogous to DNA methylation are described as being repressed rather than silenced (28, 73). Interestingly, histone modifications in PGCs and ESCs contribute to the transcriptionally permissive environment characteristic of these cell types and the reductions in DNA methylation they experience. For example, global loss of repressive H3K9 methylation marks are an essential step in epigenetic reprogramming in PGCs and induced pluripotent stem cells (74, 75).

Throughout development, most histone modifications remain dynamic as various genes are switched on and off. However, some loci, including some TE loci, appear to have a stable repertoire of histone modifications (73, 76, 77). These modifications, like DNA methylation, may be due to targeted mechanisms. As a result, TEs located next to the promoter of a gene can affect the epigenetic regulation of that gene's promoter. Therefore, new TE insertions are able to change the regulation of a gene.

TE DNA sequences are more than just repressors

A large body of evidence shows that TEs can cause large changes to gene regulatory networks. However, not all of these involve transcriptional silencing.

One of the ways TEs alter regulatory networks is through the binding of transcription factors (TFs). Using chromatin immunoprecipitation (ChIP), Bourque et al. (78) showed that several TFs had binding sites within specific TE families. Additionally, TEs with a particular TFBS were more likely to be adjacent to genes regulated by that TF than genes not regulated by that TF. One of the TFs analyzed was estrogen receptor 1 (ESR1) and was bound to MIR elements and ERV-like elements. Moreover, subsequent analyses showed these elements also harbored TFBS motifs for ESR1 co-regulators (79), thereby strengthening the idea that TFBSs from TEs affect gene expression, as the control of gene expression usually requires binding of multiple TFs (80). Further implications of combinatorial TF binding patterns found in TEs have also been linked to the evolution of particular traits. For example, MER20 is a eutherian-specific TE and is located within 200 bp of 13% of the genes associated with pregnancy in mammals (81). Of 21 randomly selected MER20s, 14 were shown to bind combinations of TFs associated with insulator activity and four were shown to bind combinations of TFs with repressor functions.

Recently, species comparisons have yielded even further insight into how TEs are able to alter regulatory networks through changes in TFBS repertoire. Schmidt et al. (82) showed that expansion of CCCTC binding factor (CTCF) binding sites in various mammalian lineages was likely due to TE expansion. CTCF is a zinc-finger protein that is able to bind DNA at a highly conserved DNA binding motif and is involved in looping DNA in long-range interactions (82–84). ChIP sequencing (ChIP-seq) characterization of CTCF binding sites in five mammalian species: human, macaque, mouse, rat, and dog showed a shared core of approximately 5000 CTCF binding sites. However, there

were also large numbers of species-specific binding sites, and many of the species-specific binding sites in mouse, rat, and dog mapped to lineage-specific TEs (both shared and unique SINE B2 elements in mouse and rat and SINEC cf elements in dog) (82). Like many of the combinations of TFs that bind to MER20s, CTCF is also a known insulator protein. Insulator proteins cause changes in gene regulation by creating chromatin boundaries. CTCF is also sensitive to methylation, and this raises questions about the extent to which TEs are transcriptionally silenced and their ability to potentially escape transcriptional silencing (85, 86). It is clear that TF binding of TEs supports a role for TEs as a potent evolutionary force in mammals. However, it is likely that binding of TFs to binding sites embedded within TEs also alters the epigenetic landscape at the TE locus.

Instrumental in the discovery of the regulatory potential of MER20s was that MER20s were enriched for chromatin marks associated with insulator activity (81). This approach has also been used in identifying the regulatory potential of other TEs in human. For example, Xie et al. (87) analyzed genome-wide methylation patterns and found that LFSINE and LTR77 TE families were differentially methylated in various tissues. Both TE families were also associated with gene expression in a tissue-specific manner and had histone modifications representative of enhancers. These findings show that epigenetic regulation is not only involved in silencing the activity of TEs but also allows them to function as enhancers or insulator elements. We can therefore say that some epigenetic regulatory mechanisms override TE-silencing mechanisms, making it more likely that TEs that contain certain TFBS are able to effectively replicate within the genome.

Transcriptional epigenetic silencing may be a powerful driver of evolution

TEs have contributed significantly to mammalian evolution in a variety of ways. The silencing of newly inserted TEs may result in an epigenetic change at a particular locus, which could therefore result in large changes in nearby gene expression, thereby altering phenotypes subject to selection.

In plant systems, Hollister et al. (88) have established that TEs contribute to an epigenetic variation that results in differences in gene expression. However, although this has not been validated in mammals, many of the components that silence TEs in plants are conserved in mammals. Comparisons between *Arabidopsis thaliana* and *Arabidopsis lyrata* revealed that sRNA-targeted TEs were associated with reduced gene expression in both species and differences in gene expression between orthologues. In addition, it was reported that changes in gene expression due to TE silencing had deleterious effects resulting in the accumulation of silenced TEs in gene-poor regions (89). This result illustrates the degree by which gene expression can be altered through silencing of TE insertions. It is important to note that some eukaryotic mechanisms responsible for silencing TEs *via* sRNAs consist of largely conserved components. Plants, fungi, and animals all use sRNAs that are cleaved by DICER-type proteins and are then bound to Argonaut proteins that then either target transcripts for RNAi or target the appropriate DNA for DNA methylation (90) (Figure 3). It is therefore likely that the observations in plants will be replicated in mammals.

Although it has not yet been shown on a genome-wide scale how TE-associated epigenetic silencing mechanisms affect gene expression in mammals, epigenomes in several mammals have been mapped. Xiao et al. compared the epigenomes of pig, mouse, and human to gain further insight into the evolution of genome-wide epigenetic regulation. Results showed that the correlations between epigenetic and gene expression conservation were higher than the correlations between sequence and gene expression conservation (91). This approach reveals that the main driver of mammalian transcriptome evolution may in fact be changes to the epigenome rather than changes in DNA sequence. It is worth noting that while patterns of epigenetic chromatin modifications may differ between mammalian species, the mechanisms that regulate them are conserved (92). For instance, the level of conservation associated with the stability of histone modifications indicates regulation of histone modifications by conserved mechanisms (93). This means that species-specific TE families can cause the same kinds of epigenetic changes in different species. Epigenetic changes resulting from heritable TE insertions can alter gene expression and hence phenotype. Therefore, TEs and their associated silencing mechanisms may have exerted significant influence on the evolution of the mammalian transcriptome.

Expert opinion

The total impact of epigenetic regulation of TEs on mammalian evolution remains largely unexplored. TEs are a major component of genome architecture, and the extent of their impact can be vast. Comparative studies involving

TEs remain a challenge due to the complexity of analyzing many closely related sequences and the economic cost of generating transcriptome and epigenome data. However, by analyzing the genomic distribution of particular TE families and developing new techniques that can compare these distributions across different species, we may be able to better understand the impact of TEs on mammalian evolution. This kind of analysis merged with transcriptome, and epigenome data will help develop a deeper understanding of the evolutionary outcomes of mammalian genomes and TE families in regard to the epigenetic mechanisms that control TE mobilization.

Highlights

- TE families are classified based on a number of criteria and have discernible features. However, an understanding of TE insertion preferences remains elusive due to divergent genomic landscapes.
- Most hypotheses aimed at reconciling the distribution of TEs use a negative selection viewpoint, that is to say, TE insertions accumulate in areas where they would be least harmful.
- The role of epigenetics in regard to TEs is largely believed to be one of defending the genome against TEs. However, some findings show that epigenetic regulation of TEs may contribute to the control of gene expression.
- Instances in which TEs provide a binding site for a DNA methylation-sensitive TF may provide TEs with an opportunity to escape transcriptional silencing.
- Several mechanisms are believed to be involved in TE silencing. However, some of these mechanisms appear to only target specific TE families.
- sRNA-mediated silencing of TEs can provide a mechanism by which TEs can alter the epigenome and pass those alterations on to the next generation. However, this has not yet been confirmed in mammals.
- Comparative studies involving epigenetics are extremely scarce because of their expense. Despite this, strong correlations have been observed between conservation in epigenomes and conservation in transcriptomes.

Outlook

AS more and more genome data become available as a result of better sequencing technologies, our understanding of the nature of genetic regulation and genome architecture will improve. One of the current bottlenecks is that both DNA and RNA sequencing analysis require assembly of short reads, usually between <100 bp and 1 kb. Because TEs are often longer than reads, it is often impossible to assemble reads from TEs accurately. Fortunately, this problem will be eliminated with the advent of sequencing technologies that use longer reads (94) such as nanopores that have the ability to read single molecules and produce reads longer than 10 kb (95).

Acknowledgments: Thanks to Dan Kortschak for helpful discussions, advice, and criticism.

References

1. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet 2009; 10: 691–703.
2. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011; 7: e1002384.
3. George CM, Alani E. Multiple cellular mechanisms prevent chromosomal rearrangements involving repetitive DNA. Crit Rev Biochem Mol Biol 2012; 47: 297–313.
4. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science 2012; 338: 758–67.
5. Robberecht C, Voet T, Esteki MZ, Nowakowska BA, Vermeesch JR. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. Genome Res 2013; 23: 411–8.
6. Hancks DC, Kazazian HH Jr. Active human retrotransposons: variation and disease. Curr Opin Genet Dev 2012; 22: 191–203.
7. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. Annu Rev Genom Hum G 2011; 12: 187–215.
8. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 2004; 431: 96–9.
9. Peng JC, Karpen GH. H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nat Cell Biol 2007; 9: 25–35.
10. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, Lorincz MC. DNA ethylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell 2011; 8: 676–87.
11. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 2008; 31: 785–99.
12. Rakyan VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KVK, Whitelaw E. Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. Proc Natl Acad Sci USA 2003; 100: 2538–43.

13. Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, Reik W. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis 2003; 35: 88–93.
14. Daxinger L, Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nat Rev Genet 2012; 13: 153–62.
15. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 2009; 448: 115–23.
16. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 2008; 9: 411–2; author reply 4.
17. Ohshima K, Okada N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res 2005; 110: 475–90.
18. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 2007; 8: 241–59.
19. Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 2002; 12: 1483–95.
20. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
21. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res 2005; 110: 117–23.
22. Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. Genet Res 2003; 82: 1–18.
23. Simons C, Pheasant M, Makunin IV, Mattick JS. Transposon-free regions in mammalian genomes. Genome Res 2006; 16: 164–72.
24. Adelson DL, Raison JM, Edgar RC. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci USA 2009; 106: 12855–60.
25. Adelson DL, Raison JM, Garber M, Edgar RC. Interspersed repeats in the horse (Equus caballus); spatial correlations highlight conserved chromosomal domains. Anim Genet 2010; 41: 91–9.
26. Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol 2012; 13: R45.
27. Castaneda J, Genzor P, Bortvin A. piRNAs, transposon silencing, and germline genome integrity. Mutat Res 2011; 714: 95–104.
28. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet 2013; 14: 204–20.
29. Hata K, Okano M, Lei H, Li E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 2002; 129: 1983–93.
30. Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, Kohara Y, Okano M, Li E, Nozaki M, Sasaki H. Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum Mol Genet 2007; 16: 2272–80.
31. Olovnikov I, Aravin AA, Fejes Toth K. Small RNA in the nucleus: the RNA-chromatin ping-pong. Curr Opin Genet Dev 2012; 22: 164–71.
32. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 2013; 27: 390–9.
33. Kelleher ES, Barbash DA. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. Mol Biol Evol 2013; 30: 1816–29.
34. Yamamoto Y, Watanabe T, Hoki Y, Shirane K, Li YF, Ichiiyanagi K, Kuramochi-Miyagawa S, Toyoda A, Fujiyama A, Oginuma M, Suzuki H, Sado T, Nakano T, Sasaki H. Targeted gene silencing in mouse germ cells by insertion of a homologous DNA into a piRNA generating locus. Genome Res 2013; 23: 292–9.

35. Hu Q, Rosenfeld MG. Epigenetic regulation of human embryonic stem cells. Front Genet 2012; 3: 238.
36. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA. Human embryonic stem cells as a model for studying epigenetic regulation during early development. Cell Cycle 2005; 4: 1323–6.
37. Maniataki E, Mourelatos Z. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. Genes Dev 2005; 19: 2979–90.
38. Hammond SM. Dicing and slicing: the core machinery of the RNA interference pathway. FEBS Lett 2005; 579: 5822–9.
39. Faulkner GJ. Retrotransposon silencing during embryogenesis: dicer cuts in LINE. PLoS Genet 2013; 9: e1003944.
40. Ciaudo C, Jay F, Okamoto I, Chen CJ, Sarazin A, Servant N, Barillot E, Heard E, Voinnet O. RNAi-dependent and independent control of LINE1 accumulation and mobility in mouse embryonic stem cells. PLoS Genet 2013; 9: e1003791.
41. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. Processing of primary microRNAs by the Microprocessor complex. Nature 2004; 432: 231–5.
42. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R. The Microprocessor complex mediates the genesis of microRNAs. Nature 2004; 432: 235–40.
43. Heras SR, Macias S, Plass M, Fernandez N, Cano D, Eyras E, Garcia-Perez JL, Cáceres JF. The Microprocessor controls the activity of mammalian retrotransposons. Nat Struct Mol Biol 2013; 20: 1173–81.
44. Meister G, Landthaler M, Peters L, Chen PY, Urlaub H, Luhrmann R, Tuschl T. Identification of novel argonaute-associated proteins. Current Biology 2005; 15: 2149–55.
45. Goodier JL, Cheung LE, Kazazian HH Jr. MOV10 RNA helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 2012; 8: e1002941.
46. Chendrimada TP, Finn KJ, Ji XJ, Baillat D, Gregory RI, Liebhaber SA, Pasquinelli AE, Shiekhattar R. MicroRNA silencing through RISC recruitment of eIF6. Nature 2007; 447: 823–U1.
47. Zheng K, Xiol J, Reuter M, Eckardt S, Leu NA, McLaughlin KJ, Stark A, Sachidanandam R, Pillai RS, Wang PJ. Mouse MOV10L1 associates with Piwi proteins and is an essential component of the Piwi-interacting RNA (piRNA) pathway. Proc Natl Acad Sci USA 2010; 107: 11841–6.
48. Frost RJA, Hamra FK, Richardson JA, Qi XX, Bassel-Duby R, Olson EN. MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs. Proc Natl Acad Sci USA 2010; 107: 11847–52.
49. Luo K, Wang T, Liu BD, Tian CJ, Xiao ZX, Kappes J, Yu XF. Cytidine deaminases APOBEC3G and APOBEC3F interact with human immunodeficiency virus type 1 integrase and inhibit proviral DNA formation. J Virol 2007; 81: 7238–48.
50. Izumi T, Burdick R, Shigemi M, Plisov S, Hu WS, Pathak VK. Mov10 and APOBEC3G localization to processing bodies is not required for virion incorporation and antiviral activity. J Virol 2013; 87: 11047–62.
51. Chiu YL, Greene WC. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annu Rev Immunol 2008; 26: 317–53.
52. Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J 2005; 24: 800–12.
53. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 2007; 8: 272–85.
54. Tachibana M, Matsumura Y, Fukuda M, Kimura H, Shinkai Y. G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. EMBO J 2008; 27: 2681–90.
55. Maksakova IA, Thompson PJ, Goyal P, Jones SJ, Singh PB, Karimi MM, Lorincz MC. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenet Chromatin 2013; 6: 15.
56. Wolf G, Nielsen AL, Mikkelsen JG, Pedersen FS. Epigenetic marking and repression of porcine endogenous retroviruses. J Gen Virol 2013; 94: 960–70.
57. Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla ME. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. Nat Struct Mol Biol 2013; 20: 332–8.
58. Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics 2008; 178: 2169–77.
59. Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci USA 2000; 97: 5237–42.
60. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev 2011; 25: 1010–22.
61. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 1999; 99: 247–57.
62. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 1992; 69: 915–26.
63. Jackson M, Krassowska A, Gilbert N, Chevassut T, Forrester L, Ansell J, Ramsahoye B. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Mol Cell Biol 2004; 24: 8862–71.
64. Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 2007; 317: 1760–4.
65. Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature 2008; 455: 818–21.
66. Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, Arrowsmith CH, Dhe-Paganon S. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature 2008; 455: 822–5.
67. Felsenfeld G, Groudine M. Controlling the double helix. Nature 2003; 421: 448–53.
68. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 2002; 298: 1039–43.
69. Nakayama J, Rice JC, Strahl BD, Allis CD, Grewal SI. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science 2001; 292: 110–3.
70. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011; 473: 43–U52.

71. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009; 459: 108–12.
72. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 2010; 107: 21931–6.
73. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, Shishkin AA, Issner R, Zhang X, Coyne M, Fostel JL, Holmes L, Meldrim J, Guttman M, Epstein C, Park H, Kohlbacher O, Rinn J, Gnirke A, Lander ES, Bernstein BE, Meissner A. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 2013; 153: 1149–63.
74. Hajkova P, Ancelin K, Waldmann T, Lacoste N, Lange UC, Cesari F, Lee C, Almouzni G, Schneider R, Surani MA. Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature 2008; 452: 877–81.
75. Chen J, Liu H, Liu J, Qi J, Wei B, Yang J, Liang H, Chen Y, Chen J, Wu Y, Guo L, Zhu J, Zhao X, Peng T, Zhang Y, Chen S, Li X, Li D, Wang T, Pei D. H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nat Genet 2013; 45: 34–42.
76. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen CA, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung WY, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 2013; 153: 1134–48.
77. Yu P, Xiao S, Xin X, Song CX, Huang W, McDee D, Tanaka T, Wang T, He C, Zhong S. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res 2013; 23: 352–64.
78. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 2008; 18: 1752–62.
79. Testori A, Caizzi L, Cutrupi S, Friard O, De Bortoli M, Cora' D, Caselle M. The role of transposable elements in shaping the combinatorial interaction of transcription factors. BMC Genomics 2012; 13: 400.
80. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, Flicek P, Odom DT. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 2013; 154: 530–40.
81. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 2011; 43: 1154–9.
82. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 2012; 148: 335–48.
83. Merkenschlager M, Odom DT. CTCF and Cohesin: linking gene regulatory elements with their targets. Cell 2013; 152: 1285–97.
84. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev 2006; 20: 2349–54.
85. Lee DH, Singh P, Tsai SY, Oates N, Spalla A, Spalla C, Brown L, Rivas G, Larson G, Rauch TA, Pfeifer GP, Szabó PE. CTCF-dependent chromatin bias constitutes transient epigenetic memory of the mother at the H19-Igf2 imprinting control region in prospermatogonia. PLoS Genet 2010; 6: e1001224.
86. Rand E, Ben-Porath I, Keshet I, Cedar H. CTCF elements direct allele-specific undermethylation at the imprinted H19 locus. Curr Biol 2004; 14: 1007–12.
87. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecek T, Weiss A, O'Geen H, Farnham PJ, Madden PA, Mungall AJ, Tam A, Kamoh B, Cho S, Moore R, Hirst M, Marra MA, Costello JF, Wang T. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 2013; 45: 836–41.
88. Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci USA 2011; 108: 2322–7.
89. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 2009; 19: 1419–28.
90. Girard A, Hannon GJ. Conserved themes in small-RNA-mediated transposon control. Trends Cell Biol 2008; 18: 136–48.
91. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen CC, Musselman M, Xie M, West FD, Lewin HA, Wang T, Zhong S. Comparative epigenomic annotation of regulatory DNA. Cell 2012; 149: 1381–92.
92. Strahl BD, Allis CD. The language of covalent histone modifications. Nature 2000; 403: 41–5.
93. Woo YH, Li WH. Evolutionary conservation of histone modifications in mammals. Mol Biol Evol 2012; 29: 1757–67.
94. Spirov AV, Holloway DM. Modeling the evolution of gene regulatory networks for spatial patterning in embryo development. Procedia Comput Sci 2013; 18: 10.1016/j.procs.2013.05.303.
95. Spirov A, Holloway D. Using evolutionary computations to understand the design and evolution of gene and cell regulatory networks. Methods 2013; 62: 39–55.

Reuben Buckley is a PhD student in the laboratory of David Adelson. Reuben completed his First Class Honours degree in 2013 under the supervision of David Adelson. Reuben's Honours research project focussed on the characterisation of genomic regions based on their repeat content and evolutionary conservation, with the aim of identifying ancestral genomic territories.

David Adelson is Professor and Chair of Bioinformatics and Computational Genetics at the University of Adelaide. Dave has been a member of several livestock genome projects, including the cow, horse and sheep. Dave pioneered the de novo identification of repeats in mammalian genome sequences in the cow and horse genome projects. David Adelson's laboratory continues to work on interspersed DNA repeats, particularly with respect to retrotransposon evolution and horizontal transfer in higher organisms.

Computational approaches for mammalian genome evolution

Mammalian genomes are complex and highly dynamic structures, they consist of many different components that interact in multiple ways. Understanding how mammalian genomes evolve while maintaining their integrity is a central question in the field of biology. Throughout this introduction I have focused specifically on how various retrotransposon types have impacted on the evolution of gene regulation. However, further investigation of their evolutionary impact requires the development of novel computational and comparative genomic approaches. Here, I briefly discuss current approaches that have been used to investigate mammalian genome evolution through retrotransposition and the important obstacles that remain in the field.

Perhaps the biggest obstacle to understanding how retrotransposons affect genome evolution is the identification and annotation of retrotransposons themselves. Approaches for performing this task split into two broad categories; library based identification and *ab initio* identification. By far the most common approach used for retrotransposon identification is library based identification, which is used by the tools *Censor* and *RepeatMasker* (Kohany et al. 2006; Smit 2004). Genomic retrotransposon sequences are identified based on sequence similarity to a known retrotransposon sequence stored in a user defined library. While Censor and *RepeatMasker* have been extremely effective in annotating the repetitive content of well studied genomes such as human and mouse, their broader applications are extremely limited. Since, the library based approach can only identify known retrotransposons that are present within the users library, unknown species-specific retrotransposons in less well studied species remain undiscovered. Alternatively, tools such as *krishna* and *RepeatScout* use an *ab initio* approach for identification of retrotransposons (Price et al. 2005; Kortschak and Adelson 2014). For example, *krishna* performs a self alignment on a species genome and identifies regions that simply appear in the genome more than once. While this approach requires no *a priori* knowledge about repetitive elements such as retrotransposons, it can be quite computationally intensive for genomes as large as those found in mammals (Zeng et al. 2017). Due to this limitation, difficulties arise when trying to identify ancestral families of retrotransposons that tend to have low levels of sequence similarity between members of the same family. This is because reducing the sequence similarity threshold for repeat identification results in a large increase in search space that ultimately causes higher memory usage and longer running times. In many cases retrotransposons are identified using a combination of both *ab initio* and library based approaches, such as with the comprehensive *ab initio* repeat pipeline (CARP) (Zeng et al. 2017). In CARP, repetitive elements initially identified using *krishna* are compiled into a library of consensus

sequences that contains known repetitive elements as well. This library is then used to identify and annotate the repetitive element landscape of a species genome, capturing both ancestral and species-specific retrotransposons.

Another important way to analyse genome evolution is to use phylogenetic approaches. The field of phylogenetics is well established and contains a variety of sophisticated models that can be applied to DNA sequence data to infer evolutionary relationships between species [reviewed in (Yang and Rannala 2012)]. The goal of phylogenetic analysis is to produce a phylogenetic tree where lineage divergences are represented as bifurcations and evolutionary rates are represented by branch lengths. Phylogenetic trees are constructed from a multiple alignment of ancestrally related DNA sequences, where relationships between sequences can be inferred using various models for DNA substitution rates. Despite their usefulness for studying evolution, classic phylogenetic approaches are most suited to single gene analyses rather than genome level analyses. One reason is that various parameters such as substitution rates are assumed to be constant throughout the sequence under analysis. Across the entirety of the genome this is rarely the case as there are many types of DNA features which happen to be under various levels of selection pressure. Another reason is that sequences undergoing phylogenetic analysis must be similar enough to carry out a multiple alignment. This is difficult with mammalian genomes as they have accumulated a large amount of insertions, deletions and rearrangements often spanning several kb (Pevzner and Tesler 2003). To overcome these challenges, classical phylogenetic approaches have been modified to handle the complexities of genome scale sequence data in a variety of ways. For example, the *multiz* tool creates genome-wide multiple alignments in the presence of large indels and genomic rearrangements by aligning fragments of multiple species genomes to a single reference (Blanchette et al. 2004). Following this, in neutrally evolving sites, such as fourfold degenerate sites, it is possible to calculate genome-wide substitution rates and background nucleotide frequencies to build a phylogenetic tree. Moreover, genome-wide phylogenetic information generated from known neutrally evolving sites can then be applied to the whole genome using the programs *phyloP* and *phastCons* (Pollard et al. 2010; Siepel et al. 2005). Both of these programs identify nucleotide positions in the reference where rates of evolution vary significantly from expectation. *PhyloP* uses several different methods to identify individual sites that have either experienced evolutionary constraint or evolutionary acceleration, where *phastCons* instead identifies regions of constraint. *PhastCons* is able to do this through the use of a hidden markov model, where the probability of a specific site belonging to a particular state is dependant on the state of the previous site (Siepel and Haussler 2005). One of the outputs from *phastCons* is a series of intervals that represent regions of high evolutionary constrain

known as conserved elements, which can be compared in multiple ways to assess the evolutionary importance of different kinds of genomic features. Importantly, by combining and adjusting different approaches it is possible to leverage the limitations of some methods against the advantages of others and identify new aspects of genome biology and evolution.

The identification of conserved elements introduces another important approach for analysing genome evolution, that is feature based analysis. In evolutionary genomics, feature based analysis involves identifying evolutionary significant genomic regions and measuring their associated feature statistics such as genome distribution, size range and frequency (Quinlan 2014). The benefits of feature analysis are that complex evolutionary and biological phenomena are reduced to a simple set of genomic intervals. Genomic features regularly analysed include retrotransposons, open chromatin sites, histone modifications, conserved elements, exons, introns, and transcription factor binding sites. Most important in an evolutionary context is that features can be mapped across species using the *liftOver* tool, making it possible to determine if certain genomic features and their activity is evolutionarily preserved (Hinrichs et al. 2006). Regardless, the accumulation of retrotransposons still creates a unique problem for mapping features across species, as species-specific retrotransposons cause gaps in genome-wide alignments. For example, features spanning a retrotransposon insertion in a reference genome may not map across to a query genome. While for most cases this is not a significant issue as retrotransposon insertions mostly represent lineage-specific evolutionary events, occasionally large regions that contain many retrotransposons are discarded from the analysis. Therefore, with the currently available suite of tools it is difficult to directly investigate patterns of independent retrotransposon accumulation in different species. Throughout this thesis I discuss new approaches for mapping genomic regions across species to investigate the impact of retrotransposons on genome evolution.

Bibliography

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., et al. (2006). The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598.

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. *BMC bioinformatics*, 7(1):474.

Kortschak, R. D. and Adelson, D. L. (2014). bíogo: a simple high-performance bioinformatics toolkit for the go language. *bioRxiv*, page 005033.

Pevzner, P. and Tesler, G. (2003). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome research*, 13(1):37–45.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121.

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, 21(suppl_1):i351–i358.

Quinlan, A. R. (2014). Bedtools: the swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, pages 11–12.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

Siepel, A. and Haussler, D. (2005). Phylogenetic hidden markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer.

Smit, A. F. (2004). Repeat-masker open-3.0. *http://www. repeatmasker. org*.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–314.

Zeng, L., Kortschak, R. D., Raison, J. M., Bertozzi, T., and Adelson, D. L. (2017). Superior ab initio identification, annotation and characterisation of tes and segmental duplications from genome assemblies. *bioRxiv*.

Chapter 2

Similar evolutionary trajectories for retrotransposon accumulation in mammals

The distribution of retrotransposons and the evolutionary forces that shape them are counter-intuitive at best; two separate families that replicate using the same machinery accumulate in distinct genomic regions. In this chapter I introduce an approach for mapping retrotransposon genomic distributions across distantly related species. I applied this approach to seven non-human mammalian genomes and found that similar retrotransposon families independently accumulated in similar genomic regions. This indicated that the forces guiding retrotransposon accumulation patterns are largely conserved across species. Finally, I introduce an open chromatin-based retrotransposon insertion model that ultimately drives similar accumulation patterns in divergent species. This chapter is in the format of a manuscript that has been submitted to the journal *Genome Biology and Evolution*.

Statement of Authorship

Title of Paper	Similar evolutionary trajectories for retrotransposon accumulation in mammals
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Buckley RM, Kortschak RD, Raison JM, Adelson DL. Similar evolutionary trajectories for retrotransposon accumulation in mammals. bioRxiv. 2017 Jan 1:091652.

Principal Author

Name of Principal Author (Candidate)	Reuben M. Buckley
Contribution to the Paper	Processed data, performed analysis, prepared figures and wrote manuscript
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 27/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	David L. Adelson
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date 25/8/2017

Name of Co-Author	R. Daniel Kortschak
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date 24/8/17

Name of Co-Author	Joy M. Raison		
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript.		
Signature		Date	1/7/17

Similar evolutionary trajectories for retrotransposon accumulation in mammals

Reuben M Buckley[1], R Daniel Kortschak[1], Joy M Raison[1], David L Adelson[1,*]

1 Department of Genetics and Evolution, The University of Adelaide, North Tce, 5005, Adelaide, Australia

*** david.adelson@adelaide.edu.au**

Keywords: Transposable element, Genome Evolution, Genome Architecture

Running title: Similar evolutionary trajectories in mammals

Abstract

The factors guiding retrotransposon insertion site preference are not well understood. Different types of retrotransposons share common replication machinery and yet occupy distinct genomic domains. Autonomous long interspersed elements accumulate in gene-poor domains and their non-autonomous short interspersed elements accumulate in gene-rich domains. To determine genomic factors that contribute to this discrepancy we analysed the distribution of retrotransposons within the framework of chromosomal domains and regulatory elements. Using comparative genomics, we identified large-scale conserved patterns of retrotransposon accumulation across several mammalian genomes. Importantly, retrotransposons that were active after our sample-species diverged accumulated in orthologous regions. This suggested a similar evolutionary interaction between retrotransposon activity and conserved genome architecture across our species. In addition, we found that retrotransposons accumulated at regulatory element boundaries in open chromatin, where accumulation of particular retrotransposon types depended on insertion size and local regulatory element density. From our results, we propose a model where density and distribution of genes and regulatory elements canalise retrotransposon accumulation. Through conservation of synteny, gene regulation and nuclear organisation, mammalian genomes with dissimilar retrotransposons follow similar evolutionary trajectories.

Introduction

An understanding of the dynamics of evolutionary changes in mammalian genomes is critical for understanding the diversity of mammalian biology. Most work on mammalian molecular evolution is on protein coding genes, based on the assumed centrality of their roles and because of the lack of appropriate methods to identify the evolutionary conservation of apparently non-conserved, non-coding sequences. Consequently, this approach addresses only a tiny fraction (less than 2%) of a species' genome, leaving significant gaps in our understanding of evolutionary processes (ENCODE Project Consortium 2012; Lander et al. 2001). In this report we describe how large scale positional conservation of non-coding, repetitive DNA sheds light on the possible conservation of mechanisms of genome evolution, particularly with respect to the acquisition of new DNA sequences.

Mammalian genomes are hierarchically organised into compositionally distinct hetero- or euchromatic large structural domains (Gibcus and Dekker 2013). These domains are largely composed of mobile self-replicating non-long terminal repeat (non-LTR) retrotransposons; with Long INterspersed Elements (LINEs) in heterochromatic regions and Short INterspersed Elements (SINEs) in euchromatic regions (Medstrand et al. 2002). The predominant LINE in most mammals is the ∼6 kb long L1. In many mammal genomes, this autonomously replicating element is responsible for the mobilisation of an associated non-autonomous SINE, usually ∼300 bp long. Together, LINEs and SINEs occupy approximately 30% of the human genome (Lander et al. 2001), replicate via a well characterised RNA-mediated copy-and-paste mechanism (Cost et al. 2002) and co-evolve with host genomes (Kramerov and Vassetzky 2011; Chalopin et al. 2015; Furano et al. 2004).

The accumulation of L1s and their associated SINEs into distinct genomic regions depends on at least one of two factors. 1) Each element's insertion preference for particular genomic regions and 2) the ability of particular genomic regions to tolerate insertions. According to the current retrotransposon accumulation model, both L1s and SINEs likely share the same insertion patterns constrained by local sequence composition. Therefore, their accumulation in distinct genomic regions is a result of region specific tolerance to insertions. Because L1s are believed to have a greater capacity than SINEs to disrupt gene regulatory structures, they are evolutionarily purged from gene-rich euchromatic domains at a higher rate than SINEs. Consequently, this selection asymmetry in euchromatic gene-rich regions causes L1s to become enriched in gene-poor heterochromatic domains (Lander et al. 2001; Graham and Boissinot 2006; Gasior et al. 2007; Kvikstad and Makova 2010).

An important genomic feature, not explored in the accumulation model, is the chromatin structure that surrounds potential retrotransposon insertion sites. Retrotransposons preferentially insert into open chromatin (Cost et al. 2001; Baillie et al. 2011; Upton et al. 2015), which is usually found overlapping gene regulatory elements. As disruption of regulatory elements can often be harmful, this creates a fundamental evolutionary conflict for retrotransposons; their immediate replication may be costly to the overall fitness of the genome in which they reside. Therefore, rather than local sequence composition or tolerance to insertion alone, retrotransposon accumulation is more likely to be constrained by an interaction between retrotransposon expression, openness of chromatin, susceptibility of a particular site to alter gene regulation, and the capacity of an insertion to impact on fitness.

To investigate the relationship between retrotransposon activity and genome evolution, we began by characterising the distribution and accumulation of non-LTR retrotransposons within placental mammalian genomes. Next, we compared retrotransposon accumulation patterns in eight separate evolutionary paths by 'humanising' the repeat content (see methods) of the chimpanzee, rhesus macaque, mouse, rabbit, dog, horse and cow genomes. Finally, we analysed human retrotransposon accumulation in large hetero- and euchromatic structural domains, focusing on regions surrounding genes, exons and regulatory elements. Our results suggest that accumulation of particular retrotransposon families follows from insertion into open chromatin found adjacent to regulatory elements and depends on local gene and regulatory element density. From this we propose a refined retrotransposon accumulation model in which random insertion of retrotransposons is primarily constrained by chromatin structure rather than local sequence composition.

Materials and Methods

Within species comparisons of retrotransposon genome distributions

Retrotransposon coordinates for each species were initially identified using RepeatMasker and obtained from either the RepeatMasker website or UCSC genome browser (Table S1) (Smit et al. 1996; Rosenbloom et al. 2015). We grouped retrotransposon elements based on repeat IDs used in Giordano *et al* (Giordano et al. 2007). Retrotransposon coordinates were extracted from hg19, mm9, panTro4, rheMac3, oruCun2, equCab2, susScr2, and canFam3 assemblies. Each species genome was segmented into 1 Mb regions and the density of each retrotransposon family for each segment was calculated. Retrotransposon density of a given genome segment is equal to a segments total number of retrotransposon nucleotides divided by that segments total number of mapped nucleotides (non-N nucleotides). From this, each species was organised into an n-by-p data matrix of n genomic segments and p retrotransposon families. Genome distributions of retrotransposons were then analysed using principle component analysis (PCA) and correlation analysis. For correlation analysis, we used our genome segments to calculate Pearson's correlation coefficient between each pair-wise combination of retrotransposon families within a species.

Across species comparisons of retrotransposon genome distributions

To compare genome distributions across species, we humanised a segmented query species genome using mapping coordinates extracted from net AXT alignment files located on the UCSC genome browser (Table S1). First, poorly represented regions were removed by filtering out genome segments that fell below a minimum mapping fraction threshold (Fig. 1a). Poorly represented regions were those that contained minimal amount of sequence alignment between pairs of species, making it difficult to accurately map non-human genomic distributions of retrotransposons to the human genome. Following this, we used these mapping coordinates to match fragments of query species segments to their corresponding human segments (Fig. 1b) and the retrotransposon content of the matched query segments were humanised following equation 1 (Fig. 1c).

$$c_i^* = \frac{\sum_j c_{ij} l_j^Q / q_j}{\sum_j l_j^R / r}, \tag{1}$$

where c_{ij} is the density of retrotransposon family i in query segment j, l_j^Q is the total length of the matched fragments between query segment j and the reference segment, l_j^R is the total length of the reference segment fragments that match query segment j, q_j is the total length of the query segment j, and r is the total length of the reference segment. The result c_i^* is the humanised coverage fraction of retrotransposon family i that can now be compared to a specific reference segment. Once genomes were humanised, Pearson's correlation coefficient was used to determine the conservation between retrotransposon genomic distributions (Fig. 1d). Using the Kolmogorov-Smirnov test, we measured the effect of humanising by comparing the humanised query retrotransposon density distribution to the query filtered retrotransposon density distribution (Fig. 1e). The same was done to measure the effect of filtering by comparing the segmented human retrotransposon density distribution to the human filtered retrotransposon density distribution (Fig. 1f). Our Pearson's correlation coefficients and P-values from measuring the effects of humanising and filtering were integrated into a heatmap (Fig. 1g). This entire process was repeated at different minimum mapping fraction thresholds to optimally represent each retrotransposon families genomic distribution in a humanised genome (fig S1).

Replication timing profiles, boundaries and constitutive domains

Genome-wide replication timing data for human and mouse were initially generated as part of the ENCODE project and were obtained from UCSC genome browser (Table S2-S3) (Yue et al. 2014; ENCODE Project Consortium 2012). For human genome-wide replication timing we used Repli-Seq smoothed wavelet signals generated by the UW ENCODE group (ENCODE Project Consortium 2012), in each cell-line we calculated the mean replication timing per 1Mb genome segment. For mouse genome-wide replication timing we used Repli-Chip wave signals generated by the FSU ENCODE group (Yue et al. 2014). Since two replicates were performed on each cell-line, we first calculated each cell-lines mean genome-wide replication timing and then used this value to calculate the mean replication timing per 1Mb genome segment. By calculating mean replication timing per 1 Mb segment we were able to easily compare large-scale genome-wide replication timing patterns across cell-lines. We obtained early replication domains (ERDs), late replication domains (LRDs) and timing transition regions (TTRs) from the gene expression omnibus (accession ID GSE53984) (Table S2). Replication domains for each dataset were identified using a deep neural network hidden Markov model (Liu et al. 2016). To determine RD boundary fluctuations of retrotransposon density, we defined ERD boundaries as the boundary of a TTR adjacent to an ERD. ERD boundaries from across each sample were pooled and retrotransposon density was calculated for 50 kb intervals from regions flanking each boundary 1 Mb upstream and downstream. Expected density and standard deviation for each retrotransposon group was derived from a background distribution generated by calculating the mean of 500 randomly sampled 50 kb genomic bins within 2000 kb of each ERD boundary, replicated 10000 times. To generate replication timing profiles for our ERD boundaries, we also calculated the mean replication timing per 50 kb intervals from across each human Repli-Seq sample. To identify constitutive ERDs and LRDs (cERDs and cLRDs), ERDs and LRDs classified by Liu *et al* (Liu et al. 2016) across each cell type were evenly split into 1 kb intervals. If the classification of 12 out of 16 samples agreed across a certain 1 kb interval, we classified that region as belonging to a cERDs or cLRDs, depending the region's majority classification of the 1 kb interval.

DNase1 cluster identification and activity

DNase1 sites across 15 cell lines were found using DNase-seq and DNase-chip as part of the open chromatin synthesis dataset for ENCODE generated by Duke University's Institute for Genome Sciences & Policy, University of North Carolina at Chapel Hill, University of Texas at Austin, European Bioinformatics Institute and University of Cambridge, Department of Oncology and CR-UK Cambridge Research Institute (Table S4) (ENCODE Project Consortium 2012). Regions where P-values of contiguous base pairs were below 0.05 were identified as significant DNase1 hypersensitive sites (ENCODE Project Consortium 2012). From this we extracted significant DNase1 hypersensitive sites from each sample and pooled them. DNase1 hypersensitive sites were then merged into DNase1 clusters. Cluster activity was calculated as the number of total overlapping pooled DNase1 hypersensitive sites. We also extracted intervals between adjacent DNase1 clusters to look for enrichment of retrotransposons at DNase1 cluster boundaries.

Extraction of intergenic and intron intervals

hg19 RefSeq gene annotations obtained from UCSC genome browser were used to extract a set of introns and intergenic intervals (Table S5). RefSeq gene annotations were merged and intergenic regions were classified as regions between the start and end of merged gene models. We used the strandedness of gene model boundaries to classify adjacent intergenic region boundaries as upstream or downstream. We discarded intergenic intervals adjacent to gene models where gene boundaries were annotated as both + and − strand. Regions between adjacent RefSeq exons within a single gene model were classified as introns. Introns interrupted by exons in alternatively spliced transcripts and introns overlapped by other gene models were excluded. Upstream and downstream intron boundaries were then annotated depending on the strandedness of the gene they were extracted from.

Interval boundary density of retrotransposons

Intervals were split in half and positions were reckoned relative to the feature adjacent boundary, where the feature was either a gene, exon, or DNase1 cluster (Fig. S2). To calculate the retrotransposon density at each position, we measured the fraction of bases at each position annotated as a retrotransposon. Next, we smoothed retrotransposon densities

by calculating the mean and standard deviation of retrotransposon densities within an expanding window, where window size grows as a linear function of distance from the boundary. This made it possible to accurately compare the retrotransposon density at positions where retrotransposon insertions were sparse and density levels at each position fluctuated drastically. At positions with a high base pair density a small window was used and at positions with a low base pair density a large window was used. Expected retrotransposon density p was calculated as the total proportion of bases covered by retrotransposons across all intervals. Standard deviation at each position was calculated as $\sqrt{np(1-p)}$, where n is the total number of bases at a given position.

Interval size bias correction of retrotransposon densities

Interval boundary density is sensitive to retrotransposon insertion preferences into intervals of a certain size (Fig. S3). To determine interval size retrotransposon density bias, we grouped intervals according to size and measured the retrotransposon density of each interval size group. Retrotransposon density bias was calculated as the observed retrotransposon density of an interval size group divided by the expected retrotransposon density, where the expected retrotransposon density is the total retrotransposon density across all intervals. Next, using the intervals that contribute to the position depth at each position adjacent to feature boundaries, we calculated the mean interval size. From this we corrected retrotransposon density at each position by dividing the observed retrotransposon density by the retrotransposon density bias that corresponded with that position's mean interval size.

Software and data analysis

All statistical analyses were performed using R (R Core Team 2015) with the packages GenomicRanges (Lawrence et al. 2013) and rtracklayer (Lawrence et al. 2009). R scripts used to perform analyses can be found at:
https://github.com/AdelaideBioinfo/retrotransposonAccumulation . All cell-line information is presented in tables S6-S7.

Results

Species selection and retrotransposon classification

We selected human, chimpanzee, rhesus macaque, mouse, rabbit, dog, horse and pig as representative placental species because of their similar non-LTR retrotransposon composition (Fig. S4-S5) and phylogenetic relationships. Retrotransposon coordinates were obtained from UCSC repeat masker tables and the online repeat masker database (Rosenbloom et al. 2015; Smit et al. 1996). We grouped non-LTR retrotransposon families according to repeat type and period of activity as determined by genome-wide defragmentation (Giordano et al. 2007). Retrotransposons were placed into the following groups; new L1s, old L1s, new SINEs and ancient elements (for families in each group see Fig. S5). New L1s and new SINEs are retrotransposon families with high lineage specificity and activity, while old L1s and ancient elements (SINE MIRs and LINE L2s) are retrotransposon families shared across taxa. We measured sequence similarity within retrotransposon families as percentage mismatch from family consensus sequences (Bao et al. 2015). We found that more recent lineage-specific retrotransposon families had accumulated a lower percentage of substitutions per element than older families (Fig. S6-S13). This confirmed that our classification of retrotransposon groups agreed with ancestral and lineage-specific periods of retrotransposon activity.

Genomic distributions of retrotransposons

To analyse the large scale distribution of retrotransposons, we segmented each species genome into adjacent 1 Mb regions, tallied retrotransposon distributions, performed principal component analysis (PCA) and pairwise correlation analysis (see methods). For PCA, our results showed that retrotransposon families from the same group tended to accumulate in the same genomic regions. We found that each individual retrotransposon group was usually highly weighted in one of the two major principal components (PC1 and PC2) (Fig. 2). Depending on associations between PCs and particular retrotransposon families we identified PC1 and PC2 as either the “lineage-specific PC” or the “ancestral PC”. Along the lineage-specific PC, new SINEs and new L1s were highly weighted, where in all species new SINEs were enriched in regions with few new L1s. Alternatively, along the ancestral PC, old L1s and ancient elements were highly weighted, where in all species except mouse — where

ancient elements and old L1s were co-located — ancient elements were enriched in regions with few old L1s (Fig. 2-3a,S14). The discordance observed in mouse probably resulted from the increased genome turnover and rearrangement seen in the rodent lineage potentially disrupting the distribution of ancestral retrotransposon families (Murphy et al. 2005; Capilla et al. 2016). In addition, the genome-wide density of ancestral retrotransposons in mouse was particularly low compared to our other species (Fig. S4-S5). However, as the relationship between mouse lineage-specific new retrotransposons is maintained, this discordance does not impact on downstream analyses. These results show that most genomic context associations between retrotransposon families are conserved across our sample species.

Retrotransposon accumulation and chromatin environment

In human and mouse, LINEs and SINEs differentially associate with distinct chromatin environments (Ashida et al. 2012). To determine how our retrotransposon groups associate with chromatin accessibility, we obtained ENCODE generated human cell line Repli-Seq data and mouse cell line Repli-ChIP data from the UCSC genome browser (ENCODE Project Consortium 2012; Yue et al. 2014). Repli-Seq and Repli-CHiP both measure the timing of genome replication during S-phase, where accessible euchromatic domains replicate early and inaccessible heterochromatic domains replicate late. Across our segmented genomes, we found a high degree of covariation between genome-wide mean replication timing and lineage-specific PC scores (Fig. 3a), new SINEs associated with early replication and new L1s associated with late replication. In addition, by splitting L1s into old and new groups, we showed a strong association between replication timing and retrotransposon age that was not reported in previous analyses (Pope et al. 2014). These results are probably not specific to a particular cell line, since genome-wide replication timing patterns are mostly highly correlated across cell lines from either species (Table S8). Moreover, early and late replicating domains from various human cell lines exhibit a high degree of overlap (Fig. S15). To confirm that lineage-specific retrotransposon accumulation associates with replication timing, we analysed retrotransposon accumulation at the boundaries of previously identified replication domains (RDs) (Liu et al. 2016). We focused primarily on early replicating domain (ERD) boundaries rather than late replicating domain (LRD) boundaries because ERD boundaries mark the transition from open chromatin states to closed chromatin states

and overlap with topologically associated domain (TAD) boundaries (Pope et al. 2014). Consistent with our earlier results, significant density fluctuations at ERD boundaries were only observed for new L1s and new SINEs (Fig. 3b). Because RD timing and genomic distributions of clade-specific retrotransposons are both largely conserved across human and mouse (Ryba et al. 2010; Yaffe et al. 2010), these results suggest that the relationship between retrotransposon accumulation and RD timing may be conserved across mammals.

The genomic distribution of retrotransposons is conserved across species

Our earlier results showed that the genomic distribution of retrotransposons is similar across species (Fig. 2). To determine whether our observations resulted from retrotransposon insertion into orthologous regions, we humanised segmented genomes of non-human species. Humanisation, began with a segmented human genome, a segmented non-human mammalian genome, and a set of pairwise alignments between both species. Using the pairwise alignments we calculated the percentage of nucleotides from each human segment that aligned to a specific non-human segment and vice-versa. This made it possible to remodel the retrotransposon content of each non-human genome segment within the human genome and essentially humanise non-human mammalian genomes (Fig. 1) (see methods). To test the precision of our humanisation process, we used the Kolmogorov-Smirnov test to compare the humanised retrotransposon density distribution of a specific retrotransposon family, to the non-humanised retrotransposon density distribution of that same retrotransposon family (Fig. S1). If the Kolmogorov-Smirnov test returned a low P-value, this suggested that the humanisation process for a given retrotransposon family had a low level of precision. Therefore, to increase our precision we used a minimum mapping fraction threshold to discard genomic segments that had only had a small amount of aligning regions between each genome. The motivation behind this was that genomic segments with a small amount of aligning sequence were the ones most likely to inaccurately represent non-human retrotransposon genomic distributions when humanised. However, it is important to note that our increase in precision requires a trade-off in accuracy. By discarding genomic segments below a certain threshold we sometimes removed a significant fraction of our non-human genomes from the analysis. In addition, this approach disproportionately affected retrotransposons such as new L1s,

as they were most enriched in segments with a small amount of aligning regions between each genome (Fig.S16-S17). To overcome this, we humanised each non-human genome at minimum mapping fraction thresholds of 0, 10, 20, 30, 40 and 50 percent and recorded the percentage of the genome that remained. We found that most retrotransposon families were precisely humanised at a minimum mapping fraction threshold of 10%. In non-human species where humanisation was most precise, a minimum mapping fraction threshold of 10% resulted in greater than 90% of the human and non-human genome remaining in the analysis (Fig. 4,S18-S24). After humanising each non-human genome, we performed pairwise correlation analysis (see methods) between the genomic distributions of each humanised and human retrotransposon family. Our results showed that retrotransposon families in different species that were identified as the same group showed relatively strong correlations, suggesting that they accumulated in regions with shared common ancestry (Fig. 4,S18-S24). Next, we assessed the level of conservation of retrotransposon accumulation patterns across all of our species. For each retrotransposon group in each humanised genome, we identified the top 10% retrotransposon dense genome segments. We found that when these segments were compared with the human genome, there was a relativity high degree of overlap (Fig. 5a-b). These results suggest that lineage-specific retrotransposon accumulation may follow an ancient conserved mammalian genome architecture.

Retrotransposon insertion in open chromatin surrounding regulatory elements

Retrotransposons preferentially insert into open chromatin, yet open chromatin usually overlaps gene regulatory elements. As stated above, this creates a fundamental evolutionary conflict for retrotransposons; their immediate replication may be detrimental to the overall fitness of the genome in which they reside. To investigate retrotransposon insertion/accumulation dynamics at open chromatin regions, we analysed DNase1 hypersensitive activity across 15 cell lines in both ERDs and LRDs. DNase1 hypersensitive sites obtained from the UCSC genome browser (ENCODE Project Consortium 2012) were merged into DNase1 clusters and DNase1 clusters overlapping exons were excluded. As replication is sometimes cell type-specific we also constructed a set of constitutive ERDs and LRDs (cERDs and cLRDs) (see methods). Based on previous analyses, cERDs and cLRDs likely capture

RD states present during developmental periods of heritable retrotransposition (Rivera-Mulia et al. 2015). Our cERDs and cLRDs capture approximately 50% of the genome and contain regions representative of genome-wide intron and intergenic genome structure (Fig. S25). In both cERDs and cLRDs, we measured DNase1 cluster activity by counting the number of DNase1 peaks that overlapped each cluster. We found that DNase1 clusters in cERDs were much more active than DNase1 clusters in cLRDs (Fig. 6a). Next, we analysed retrotransposon accumulation both within and at the boundaries of DNase1 clusters. Consistent with disruption of gene regulation by retrotransposon insertion, non-ancient retrotransposon groups were depleted from DNase1 clusters (Fig. 6b). Intriguingly, ancient element density in DNase1 clusters remained relatively high, suggesting that some ancient elements may have been exapted. At DNase1 cluster boundaries after removing interval size bias (Fig. S26-S27) (see methods), retrotransposon density remained highly enriched in cERDs and close to expected levels in cLRDs (Fig. 6c). This suggests that chromatin is likely to be open at highly active cluster boundaries where insertion of retrotransposons is less likely to disrupt regulatory elements. To confirm that recent retrotransposon insertion follows open chromatin we analysed the accumulation patterns of individual human retrotransposon families that were active at different periods during primate evolution. The families we chose were *AluY*, L1HS, *AluJ* and L1MA. *AluY* and L1HS are mostly human-specific while *AluJ* and L1MA were most active in an ancestral primate (Mills et al. 2007). We found that elements from younger retrotransposon families were more enriched near DNase1 cluster boundaries than elements from older retrotransposon families from within the same retrotransposon group (Fig. S28). Collectively, these results are consistent with an interaction between retrotransposon insertion, open chromatin and regulatory activity, where insertions into open chromatin only persist if they do not interrupt regulatory elements.

Retrotransposon insertion size and regulatory element density

L1s and their associated SINEs differ in size by an order of magnitude, retrotranspose via the L1-encoded chromatin-sensitive L1ORF2P and accumulate in compositionally distinct genomic domains (Cost et al. 2001). This suggests that retrotransposon insertion size determines observed accumulation patterns. L1 and *Alu* insertions occur via target-primed reverse transcription which is initiated at the $3'$ end of each element. With L1 insertion, this

process often results in 5′ truncation, causing extensive insertion size variation and an over representation of new L1 3′ ends, not seen with *Alu* elements (Fig. 7a). When we compared insertion size variation across cERDs and cLRDs we observed that smaller new L1s were enriched in cERDs and *Alu* elements showed no RD insertion size preference (Fig. 7b). The effect of insertion size on retrotransposon accumulation was estimated by comparing insertion rates of each retrotransposon group at DNase1 cluster boundaries in cERDs and cLRDs. We found that *Alu* insertion rates at DNase1 cluster boundaries were similarly above expected levels both in cERDs and cLRDs (Fig. 7c), whereas new L1 insertion rates at DNase1 cluster boundaries were further above expected levels in cERDs than cLRDs (Fig. 7d). By comparing the insertion rate of new L1s — retrotransposons that exhibited RD specific insertion size variation — we observed a negative correlation between element insertion size and gene/regulatory element density. Thus smaller elements, such as *Alu* elements, accumulate more in cERDs than do larger elements, such as new L1s, suggesting that smaller elements are more tolerated.

Retrotransposon insertion within gene and exon structures

Regulatory element organisation is largely shaped by gene and exon/intron structure which likely impacts the retrotransposon component of genome architecture. Therefore, we analysed retrotransposons and DNase1 clusters (exon-overlapping and exon non-overlapping) at the boundaries of genes and exons. Human RefSeq gene models were obtained from the UCSC genome browser and both intergenic and intronic regions were extracted (Table S5). At gene (Fig. 8a) and exon (Fig. 8b) boundaries, we found a high density of exon overlapping DNase1 clusters and depletion of retrotransposons. This created a depleted retrotransposon boundary zone (DRBZ) specific for each retrotransposon group, a region extending from the gene or exon boundary to the point where retrotransposon levels begin to increase. The size of each DRBZ correlated with the average insertion size of each retrotransposon group, consistent with larger retrotransposons having a greater capacity to disrupt important structural and regulatory genomic features. We also found that in cERDs the 5′ gene boundary *Alu* DRBZ was larger than the 3′ gene boundary *Alu* DRBZ. This difference was associated with increased exon overlapping DNase1 cluster density at 5′ gene boundaries in cERDs (Fig. 8a), emphasising the importance of evolutionary constraints on promoter

architecture. For ancient elements, their retrotransposon density at approximately 1 kb from the 5′ gene boundary, when corrected for interval size bias, was significantly higher than expected. This increase is consistent with exaptation of ancient elements into regulatory roles (Lowe et al. 2007) (Fig. S29-S32). Moreover, the density peak corresponding to uncorrected ancient elements also overlapped with that of exon non-overlapping DNase1 clusters (Fig. 8a). Collectively, these results demonstrate the evolutionary importance of maintaining gene structure and regulation and how this in turn has canalised similar patterns of accumulation and distribution of retrotransposon families in different species over time.

Discussion

A conserved architectural framework shapes the genomic distribution of ancestral retrotransposons

The majority of divergence between our sample species has taken place over the last 100 million years. Throughout this time period many genomic rearrangements have occurred, causing a great deal of karyotypic variation. However, we found that the genomic distributions of ancestral elements remained conserved. The evolutionary forces preserving the ancestral genomic distributions of these elements remain unclear.

One suggestion is that ancestral elements play essential roles in mammalian organisms. Our results in Fig. 6b and 8a suggest that ancient elements have been exapted. Their accumulation within open chromatin sites is consistent with their roles as *cis*-regulatory element, such as MIR elements that perform as TFBSs and enhancers (Bourque et al. 2008; Jjingo et al. 2014). Similarly, L1s also carry binding motifs for DNA-binding proteins. L1 elements that were active prior to the boreoeutherian ancestor bind a wide variety of KRAB zinc-finger proteins (KZFPs), most of which have unknown functions (Imbeault et al. 2017). In terms of genome structural roles, some human MIR elements have been identified as insulators, separating open chromatin regions from closed chromatin regions (Wang et al. 2015). While these MIR insulators function independently of CTCF binding, their mechanism of action remains largely unknown. Despite this, when a human MIR insulator was inserted into the zebrafish genome it was able to maintain function (Wang et al. 2015). This suggests that MIR insulators recruit a highly conserved insulator complex and maintain insulator

function across the mammalian lineage. Collectively, these findings identified a number of examples where ancestral elements are associated with important biological roles. This may suggest that genomic distributions of ancestral elements are conserved across mammals because they play conserved biological roles across mammals. However, it is necessary to draw a distinction between evolutionary conservation of an ancient functional element and evolutionary conservation of large-scale genomic distributions of retrotransposons. This is important because for most of our sample species, ancient elements and old L1s each occupy approximately 7% of each of their genomes (Fig. S4). Compared to the 0.04% of the human genome that is comprised of transposable elements under purifying selection (Lowe et al. 2007), this suggests that the vast majority of ancestral elements may not actually play conserved roles in mammalian biology.

Rather than ancestral elements playing a conserved role in genome maintenance, their genomic distributions may instead remain conserved as a consequence of evolutionary dynamics occurring at higher order levels of genome architecture. TADs have been identified as a fundamental unit of genome structure, they are approximately 900 kb in length and contain highly self interacting regions of chromatin (Dixon et al. 2012). Despite large-scale genomic rearrangements, the boundaries between TADs have remained conserved across mammals (Dixon et al. 2012). An analysis involving rhesus macaque, dog, mouse and rabbit, identified TAD boundaries at the edge of conserved syntenic regions associating with evolutionary breakpoints between genomic rearrangements (Rudan et al. 2015). This suggests that genome rearrangements occur primarily along TAD boundaries leaving TADs themselves largely intact. Similarly, TAD architecture could also be the driving force behind the observed frequent reuse of evolutionary breakpoints throughout mammalian genome evolution (Murphy et al. 2005). Together these findings suggest that TADs form part of a conserved evolutionary framework whose boundaries are sensitive to genomic rearrangements. Therefore, the current observed genomic distributions of ancestral retrotransposons reflects mostly ancestral retrotransposons that inserted within TADs rather than at their boundaries. This is because elements that accumulated near TAD boundaries were most likely lost through recurrent genomic rearrangements and genome turnover.

Another example supporting the idea that conserved genomic distributions are shaped by a conserved architectural evolutionary framework can be found in the rodent lineage. Rodents have experienced rates of genome reshuffling two orders of magnitude greater than

other mammalian lineages (Capilla et al. 2016). This has caused rodent genomes to contain a higher number of evolutionary breakpoints, many of which are rodent-specific (Capilla et al. 2016). From our analysis we found that old L1s and ancient elements each occupied only 1% of the mouse genome (Fig. S4), with similar levels of ancient elements within the rat genome (Gibbs et al. 2004). Compared to our other species where the genomes are approximately 7% ancient elements and old L1s each (S4), rodent genomes are significantly depleted of ancestral elements. Together, these findings show a negative correlation between ancestral retrotransposon content and rate of genome rearrangements, suggesting that increased rates of genome rearrangements can strongly impact the genomic distributions of ancestral retrotransposons. In addition, the large number of rodent specific evolutionary breakpoints may explain why the genomic distribution of ancestral elements in mouse is discordant with our other species. Specifically, ancient elements and old L1s in mouse accumulated in similar regions, whereas in each of our other species ancient elements and old L1s accumulated in almost opposite regions as defined by PC1 (Fig. 2,3a).

Conserved genome architecture drives the accumulation patterns of lineage-retrotransposons

Across mammals, lineage-specific retrotransposons are responsible for the vast majority of lineage-specific DNA gain (Kapusta et al. 2017). Throughout our sample-species we found that new SINEs and new L1s independently accumulated in similar regions in different species. These results suggest there is a high degree of conservation surrounding their insertion mechanisms and genomic environments. Since, L1 conservation in mammals is well documented in the literature and our new SINE families all replicate using L1 machinery, mainly we spend this section discussing the role of conserved genome architecture (Ivancevic et al. 2016; Vassetzky and Kramerov 2013).

Earlier, we discussed the importance of TADs and how they form a fundamental component of conserved genome architecture. This same architectural framework may also shape the accumulation pattern of lineage specific retrotransposons. TAD boundaries separate the genome into regions comprised of genes that are largely regulated by a restricted set of nearby enhancers. Moreover, TADs are subject to large-scale changes in chromatin structure, where individual TADs are known to switch between open and closed chromatin states in a

cell type-specific manner (Dixon et al. 2012). One method of capturing shifts in chromatin state between TADs is to measure genome-wide replication timing (Pope et al. 2014). This is because replication timing associates with the genomes accessibility to replication machinery. Accessible regions that comprise an open chromatin structure replicate early while inaccessible regions with a closed chromatin structure replicate late. Genome-wide replication timing follows a domain-like organisation, where large contiguous regions either replicate at earlier or later stages of mitosis. Importantly, ERD boundaries directly overlap TAD boundaries, supporting the notion that TADs are also fundamental units of large-scale chromatin state organisation (Pope et al. 2014). Previously, LINE and SINE accumulation patterns were associated with TAD and RD genome architecture, where LINEs were enriched in LRDs and SINEs were enriched in ERDs (Hansen et al. 2010; Rivera-Mulia et al. 2015; Pope et al. 2014; Ashida et al. 2012). Unlike our analysis, these earlier studies decided not to separate LINEs into ancestral and lineage-specific families. Despite this difference, Fig. 3 shows that our results are consistent with earlier analyses, except for our observation that only lineage-specific retrotransposon families are associated with replication timing. Therefore, by separating L1s and SINEs according to period of activity, we observed much stronger associations between replication timing and retrotransposon accumulation than previously reported (Pope et al. 2014; Ashida et al. 2012). Since replication timing and boundaries between TADs and RDs are conserved across mammalian species (Ryba et al. 2010; Yaffe et al. 2010; Pope et al. 2014; Dixon et al. 2012), our results suggest that domain-level genome architecture likely plays a role in shaping conserved lineage-specific retrotransposon accumulation patterns.

While our species genomes are conserved at a structural level, conserved patterns of lineage-specific retrotransposon accumulation can have significant evolutionary impacts. new SINEs accumulate in ERDs which tend to be highly active gene-rich genomic regions. However, despite the fact that all of our new SINE families follow L1 mediated replication, they stem from unique origins. For example, Primate-specific *Alu* elements are derived from 7SL RNA and carnivora-specifc SINEC elements are dervided form tRNA (Quentin 1994; Coltman and Wright 1994). Due to their large-scale accumulation patterns this means that new SINEs in mammalian genomes simultaneously drive convergence in genome architecture and divergence in genome sequence composition. This is especially important because SINEs are also a large source of evolutionary innovation for gene regulation. In human, various

individual *Alu* elements have been identified as bona fide enhancers with many more believed to be proto-enhacers serving as a repertoire for birth of new enhancers (Su et al. 2014). Similarly, in dog, mouse and opossum, lineage specific SINEs carry CTCF binding sites and have driven the expansion of species-specific CTCF binding patterns (Schmidt et al. 2012).

Like new SINEs, new L1s also accumulate in similar regions in different species. However, unlike new SINEs, lineage-specific mammalian L1 elements most likely stem from a common ancestor (Furano et al. 2004). This means that individual new L1 elements in different species are more likely than species-specific SINEs to share similar sequence composition (Ivancevic et al. 2016). Therefore, LRDs, which are enriched for new L1s, may show higher levels of similarity for genome sequence composition than ERDs, which are enriched for new SINEs. Considering results from genome-wide alignments between mammals, this may be counter intuitive, mainly because the surrounding sequence in new L1 enriched regions exhibits poor sequence conservation (Fig. S16-S17). However, it is important to realise that similar sequence composition is not the same as sequence conservation itself, especially at the level of mammalian genome architecture. Sequence composition refers to the kinds of sequences in a particular region rather than the entire sequence of the region itself. For example, binding sites for the same transcription factor in different species are sometimes located in similar regions yet differ in position relative to their target genes (Kunarso et al. 2010). So while genome-wide alignments may suggest low levels of genome conservation or high levels of turnover, sequence composition within these regions remains similar and can still be indicative of conserved function. Therefore with the accumulation of new L1s after species divergence, it is likely that sequence conservation decreases at a much faster rate than compositional similarity. For new L1s enriched in similar regions in different species, this may have important functional consequences. Recently, highly conserved ancient KZFPs were discovered to bind to members of both old and new L1 families in human (Imbeault et al. 2017). This suggests that new L1s in humans may be interchangeable with old L1s and play important roles in highly conserved gene regulatory networks. Therefore, because new L1s in different species share similar sequences and their accumulation patterns are also conserved, new L1s may actively preserve ancient gene regulatory networks across the mammalian lineage.

A chromatin based model of retrotransposon accumulation

Analysis of repetitive elements in mammalian genome sequencing projects has consistently revealed that L1s accumulate in GC-poor regions and their mobilised SINEs accumulate in GC-rich regions (Lander et al. 2001; Gibbs et al. 2004; Chinwalla et al. 2002). Our results were consistent with this and showed that accumulation patterns of new SINEs and new L1s were conserved across species and corresponded with distinct genomic environments. Since these elements both replicate via the same machinery, their accumulation patterns are most likely shaped by how insertion of each element type interacts with its immediate genomic environment. The current model of retrotransposon accumulation begins with random insertion, constrained by local sequence composition, followed by immediate selection against harmful insertions (Graham and Boissinot 2006; Gasior et al. 2007; Kvikstad and Makova 2010). During early embryogenesis or in the germline, it is believed retrotransposons in individual cells randomly insert into genomic loci that contain a suitable insertion motif. Because this process is assumed to be random, new insertions can occasionally interrupt essential genes or gene regulatory structures. These insertions are usually harmful, causing the individual cell carrying them to be quickly removed from the population. This process of purifying selection prevents harmful insertions from being passed down to the next generation and plays a large role in shaping retrotransposon accumulation patterns. According to this model, because of their size difference L1s are considered to have a more harmful impact on nearby genes and gene regulatory structures than SINEs. New L1 insertion into GC-rich regions, which are also gene-rich, are more likely to cause harm than if new SINEs inserted into those same regions. Therefore, new L1s are evolutionary purged from GC rich regions causing them to become enriched in gene-poor AT-rich regions. While this model is simple, it fails to take into account the impact of chromatin structure that constrains retrotransposon insertion preference. Therefore, we decided to analyse retrotransposon accumulation at the level of large-scale chromosomal domains and fine-scale open chromatin sites.

Our results showed that lineage-specifc retrotransposons accumulated at the boundaries of open chromatin sites. This was particularly striking as it appeared to reconcile insertion into open chromatin with the risk of disrupting regulatory elements. Single cell analysis has shown somatic retrotransposition events correlate with preferable insertion into open chromatin sites or within actively expressed genes (Klawitter et al. 2016; Upton et al. 2015; Baillie

et al. 2011). However, because open chromatin usually surrounds regulatory elements these kinds of insertions can be a major cause of genetic disease (Wimmer et al. 2011). Therefore, retrotransposons accumulate in open chromatin regions where their insertion is less likely to disrupt regulatory elements. We further demonstrated the impact of retrotransposon insertion by considering element insertion size. Our results showed that shorter L1s were much more likely to insert close to open chromatin sites surrounding regulatory elements than larger L1s. This suggested that L1 insertions were much more likely than *Alu* insertions to impact on gene regulatory structures due to their larger insertion size. At this point, it should be noted that chromatin state can be highly dynamic, switching between open and closed states depending on cell type (ENCODE Project Consortium 2012). Importantly, heritable retrotransposon insertions typically occur during embryogenesis or within the germline. However, chromatin state data for these developmental stages and tissue samples was unavailable. To overcome this limitation we aggregated data from a range of biological contexts. The underlying assumption behind this strategy was that open chromatin sites found in at least one cell likely contain regulatory elements that may be reused in another cell type. By using this strategy, we increased the probability of capturing chromosomal domain structures and regulatory element sites present in embryonic and germline cell states. While our strategy may help overcome limitations regarding unavailable cell types, it is still not the definitive test of our model of retrotransposon insertion into open chromatin. Ultimately, the necessary data would require a robust cell-line with tens to hundreds of thousands of known *de novo* retrotransposon insertions complete with genome-wide chromatin state data. Under our model we would expect that *de novo* insertions would be enriched in open chromatin sites. Additionally, because of relaxation in selective pressures in cell-line, insertions would not necessarily accumulate at regulatory element boundaries like they do in reference genomes. Alternatively, if chromatin state were not a driving factor shaping initial insertion accumulation patterns, we would expect no observable increase in insertion rates at open chromatin sites. Previously, a similar approach was used in HeLa cells and proved to be very powerful in identifying the sequence context of L1 insertions and L1 mediated genomic rearrangements (Gilbert et al. 2002, 2005). This was largely because cell-lines make it possible to disentangle the confounding effects of retrotransposon activity and purifying selection at insertion sites.

An important aspect of both our refined model and the current model of retrotransposon

accumulation is the immediate evolutionary impact of retrotransposon insertions. Specifically, at what rate do embryonic and germline retrotransposition events occur and what proportion of these events escape purifying selection? Answering this question is a challenging task primarily limited by the availability of samples at the correct developmental time periods. Ideally we would require genome sequencing data from a large population of germline or embryonic cells derived from a similar genetic background. Given that data, we could identify new insertions before they have undergone selection and compare their retrotransposition rates to retrotransposition rates inferred from population data. Alternatively, retrotransposition rates have been measured in somatic cells and stem-cell lines. In hippocampal neurons and glia, L1 retrotransposition occurs at rates of 13.7 and 6.5 events per cell, where in human induced pluripotent stem cells retrotransposition rates are approximately 1 event per cell (Klawitter et al. 2016; Upton et al. 2015). In neurons, L1 insertions were enriched in neuronally expressed genes and in human induced pluripotent stem cells, L1s were found to insert near transcription start sites, disrupting the expression of some genes (Klawitter et al. 2016; Upton et al. 2015; Baillie et al. 2011). This suggests L1s are particularly active in humans, able to induce a large amount of variation and disrupt gene regulation and function. It is also important to note that the estimated L1 heritable retrotransposition rate is approximately one event per 95 to 270 births (Ewing and Kazazian 2010), suggesting that many insertions are removed from the germline cell population. For *Alu* elements this rate is much greater, *Alu* elements are estimated to undergo heritable retrotransposition at a rate of one event per 20 births (Cordaux et al. 2006). These findings support the notion that the majority of retrotransposon insertions are likely to be evolutionarily purged from the genome.

In summary, by analysing open chromatin sites, we found that 1) following preferential insertion into open chromatin domains, retrotransposons were tolerated adjacent to regulatory elements where they were less likely to cause harm; 2) element insertion size was a key factor affecting retrotransposon accumulation, where large elements accumulated in gene poor regions where they were less likely to perturb gene regulation; and 3) insertion patterns surrounding regulatory elements were persistent at the gene level. From this we propose a significant change to the current retrotransposon accumulation model; rather than random insertion constrained by local sequence composition, we propose that insertion is instead primarily constrained by local chromatin structure. Therefore, L1s and SINEs

both preferentially insert into gene/regulatory element rich euchromatic domains, where L1s with their relatively high mutational burden are quickly eliminated via purifying selection at a much higher rate than SINEs. Over time this results in an enrichment of SINEs in euchromatic domains and an enrichment of L1s in heterochromatic domains.

Conclusion

In conjunction with large scale conservation of synteny (Chowdhary et al. 1998), gene regulation (Chan et al. 2009) and the structure of RDs/TADs (Dixon et al. 2012; Ryba et al. 2010),our findings suggest that large scale positional conservation of old and new non-LTR retrotransposons results from their association with the regulatory activity of large genomic domains. Therefore we propose that similar constraints on insertion and accumulation of clade specific retrotransposons in different species can define common trajectories for genome evolution.

Additional Files

Additional file 1 — Supplementary information

Figures S1–S31, Tables S1–S6.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

R.M.B., R.D.K., J.M.R., and D.L.A. designed research; R.M.B. performed research; and R.M.B., R.D.K., and D.L.A. wrote the paper.

Acknowledgements

For reviewing our manuscript and providing helpful advice we would like to thank the following: Simon Baxter, Atma Ivancevic and Lu Zeng from the University of Adelaide; Kirsty Kitto from Queensland University of Technology; and Udaya DeSilva from Oklahoma State University.

Availability of data and materials

All data was obtained from publicly available repositories, urls can be found in supporting material (Table S1–S4). R scripts used to perform analyses can be found at https://github.com/AdelaideBioinfo/retrotransposonAccumulation.

References

Ashida, H., Asai, K., and Hamada, M. (2012). Shape-based alignment of genomic landscapes in multi-scale resolution. *Nucleic acids research*, 40(14):6435–6448.

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P. M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, 479(7374):534–537.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6:11.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11):1752–1762.

Capilla, L., Sánchez-Guillén, R. A., Farre, M., Paytuví-Gallart, A., Malinverni, R., Ventura, J., Larkin, D. M., and Ruiz-Herrera, A. (2016). Mammalian comparative genomics reveals genetic and epigenetic features associated with genome reshuffling in rodentia. *Genome Biology and Evolution*, 8(12):3703–3717.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome biology and evolution*, 7(2):567–580.

Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3):33.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton,

R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Chowdhary, B. P., Raudsepp, T., Frönicke, L., and Scherthan, H. (1998). Emerging patterns of comparative genome organization in some mammalian species as revealed by zoo-fish. *Genome research*, 8(6):577–589.

Coltman, D. W. and Wright, J. M. (1994). Can sines: a family of trna-derived retroposons specific to the superfamily canoidea. *Nucleic acids research*, 22(14):2726–2730.

Cordaux, R., Hedges, D. J., Herke, S. W., and Batzer, M. A. (2006). Estimating the retrotransposition rate of human alu elements. *Gene*, 373:134–137.

Cost, G. J., Feng, Q., Jacquier, A., and Boeke, J. D. (2002). Human l1 element target-primed reverse transcription in vitro. *The EMBO Journal*, 21(21):5899–5910.

Cost, G. J., Golding, A., Schlissel, M. S., and Boeke, J. D. (2001). Target dna chromatinization modulates nicking by l1 endonuclease. *Nucleic acids research*, 29(2):573–577.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Ewing, A. D. and Kazazian, H. H. (2010). High-throughput sequencing reveals extensive variation in human-specific l1 content in individual human genomes. *Genome research*, 20(9):1262–1270.

Furano, A. V., Duvernell, D. D., and Boissinot, S. (2004). L1 (line-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends in Genetics*, 20(1):9–14.

Gasior, S. L., Preston, G., Hedges, D. J., Gilbert, N., Moran, J. V., and Deininger, P. L. (2007). Characterization of pre-insertion loci of de novo l1 insertions. *Gene*, 390(1):190–198.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., et al. (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521.

Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782.

Gilbert, N., Lutz, S., Morrish, T. A., and Moran, J. V. (2005). Multiple fates of l1 retrotransposition intermediates in cultured human cells. *Molecular and cellular biology*, 25(17):7780–7795.

Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). Genomic deletions created upon line-1 retrotransposition. *Cell*, 110(3):315–325.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., and Warburton, P. E. (2007). Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput Biol*, 3(7):e137.

Graham, T. and Boissinot, S. (2006). The genomic distribution of L1 elements: the role of insertion bias and natural selection. *BioMed Research International*, 2006(1):75327.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). Krab zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554.

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., and Adelson, D. L. (2016). Lines between species: Evolutionary dynamics of line-1 retrotransposons across the eukaryotic tree of life. *Genome Biology and Evolution*, 8(11):3301–3322.

Jjingo, D., Conley, A. B., Wang, J., Mariño-Ramírez, L., Lunyak, V. V., and Jordan, I. K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA*, 5:14.

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8):E1460–E1469.

Klawitter, S., Fuchs, N. V., Upton, K. R., Muñoz-Lopez, M., Shukla, R., Wang, J., Garcia-Cañadas, M., Lopez-Ruiz, C., Gerhardt, D. J., Sebe, A., et al. (2016). Reprogramming

triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nature Commun*, 7:10286.

Kramerov, D. and Vassetzky, N. (2011). Origin and evolution of sines in eukaryotic genomes. *Heredity*, 107(6):487–495.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics*, 42(7):631–634.

Kvikstad, E. M. and Makova, K. D. (2010). The (r) evolution of sine versus line distributions in primate genomes: sex chromosomes are important. *Genome research*, 20(5):600–613.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8):e1003118.

Liu, F., Ren, C., Li, H., Zhou, P., Bo, X., and Shu, W. (2016). De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics*, 32(5):641–649.

Lowe, C. B., Bejerano, G., and Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, 104(19):8005–8010.

Medstrand, P., Van De Lagemaat, L. N., and Mager, D. L. (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome research*, 12(10):1483–1495.

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in genetics*, 23(4):183–191.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405.

Quentin, Y. (1994). Emergence of master sequences in families of retroposons derived from 7sl rna. *Genetica*, 93(1-3):203–215.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rivera-Mulia, J. C., Buckley, Q., Sasaki, T., Zimmerman, J., Didier, R. A., Nazor, K., Loring, J. F., Lian, Z., Weissman, S., Robins, A. J., et al. (2015). Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome research*, 25(8):1091–1103.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., et al. (2015). The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681.

Rudan, M. V., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. (2015). Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell reports*, 10(8):1297–1309.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T. C., Robins, A. J., Dalton, S., and Gilbert, D. M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20(6):761–770.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and ctcf binding in multiple mammalian lineages. *Cell*, 148(1):335–348.

Smit, A. F., Hubley, R., and Green, P. (1996). Repeatmasker open-3.0.

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.-D. J. (2014). Evolution of alu elements toward enhancers. *Cell reports*, 7(2):376–385.

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., Ewing, A. D., Salvador-Palomeque, C., van der Knaap, M. S., Brennan, P. M., et al. (2015). Ubiquitous l1 mosaicism in hippocampal neurons. *Cell*, 161(2):228–239.

Vassetzky, N. S. and Kramerov, D. A. (2013). Sinebase: a database and tool for sine analysis. *Nucleic acids research*, 41(D1):D83–D89.

Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., Lee, E., Gómez-Skarmeta, J. L., Montoliu, L., Lunyak, V. V., et al. (2015). Mir retrotransposon sequences provide insulators to the human genome. *Proceedings of the National Academy of Sciences*, 112(32):E4428–E4437.

Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The nf1 gene contains hotspots for l1 endonuclease-dependent de novo insertion. *PLoS Genet*, 7(11):e1002371.

Yaffe, E., Farkash-Amar, S., Polten, A., Yakhini, Z., Tanay, A., and Simon, I. (2010). Comparative analysis of dna replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet*, 6(7):e1001011.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527):355–364.

Figures

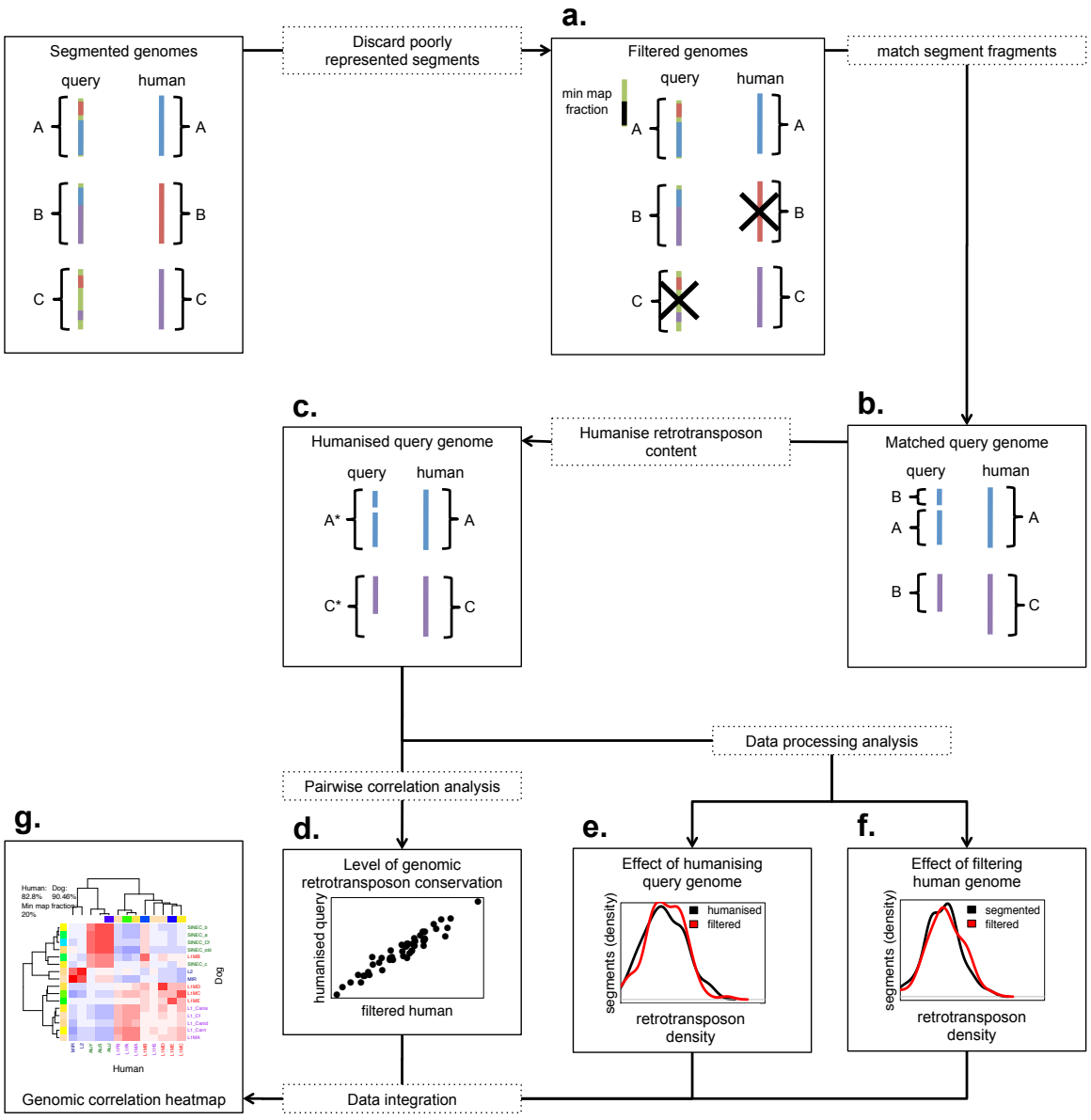


Figure 1. Overview of humanising retrotransposon distributions. **a,** Genomes are segmented and filtered according to a minimum mapping fraction threshold, removing poorly represented segments from both species. The black X shows which segments were not able to reach the minimum mapping fraction threshold. **b,** Fragments of query species' genome segments are matched to their corresponding human genome segments using genome alignments. **c,** Query species genomes are humanised following equation 1. **d,** Pairwise genomic correlations are measured between each humanised retrotransposon family and each human retrotransposon family. **e,** The effect of humanising on retrotransposon density distributions is measured by performing a Kolmogorov-Smirnov test between the humanised query retrotransposon density distribution and the filtered query retrotransposon density distribution. **f,** The effect of filtering on retrotransposon density distributions is measured by performing a Kolmogorov-Smirnov test between the segmented human retrotransposon density distribution and the filtered human retrotransposon density distribution. **g,** The pairwise correlation analysis results and the P-values from the Kolmogorov-Smirnov tests are integrated into heatmaps (Fig. 4,S18-S22) that compare the genomic relationships of retrotransposons between species.

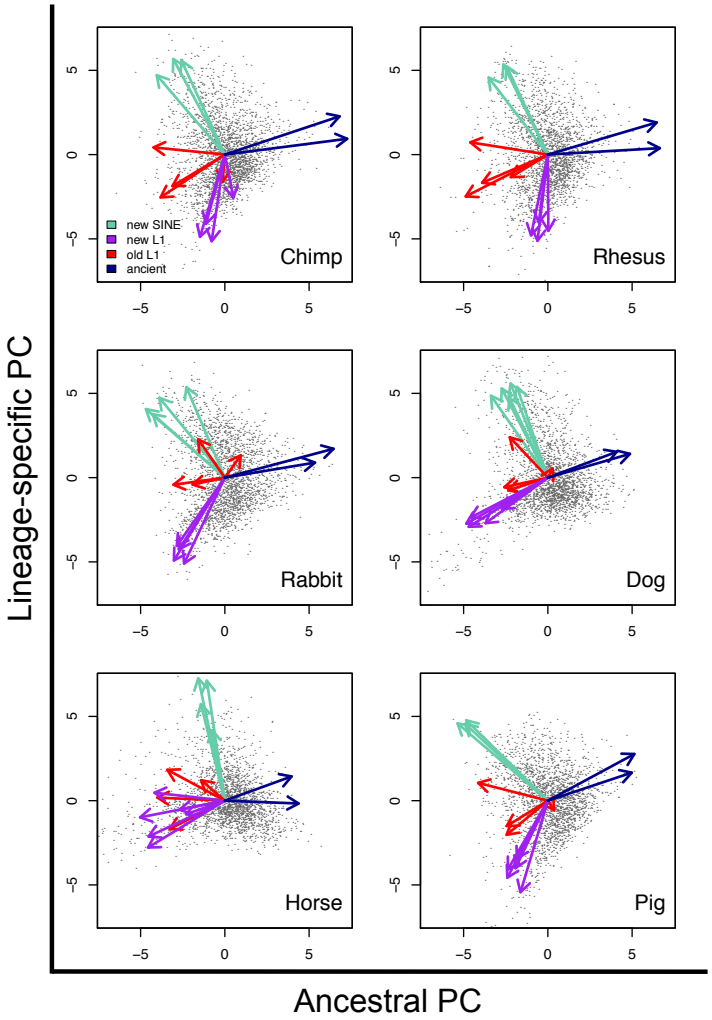


Figure 2. Similar genomic distributions of retrotransposons across mammals. Principal Component 1 and Principal Component 2 of non-human and non-mouse genome retrotransposon content, each vector loading has been coloured according to the retrotransposon group it represents. Principal components have been renamed according to the retrotransposon group whose variance they principally account for.

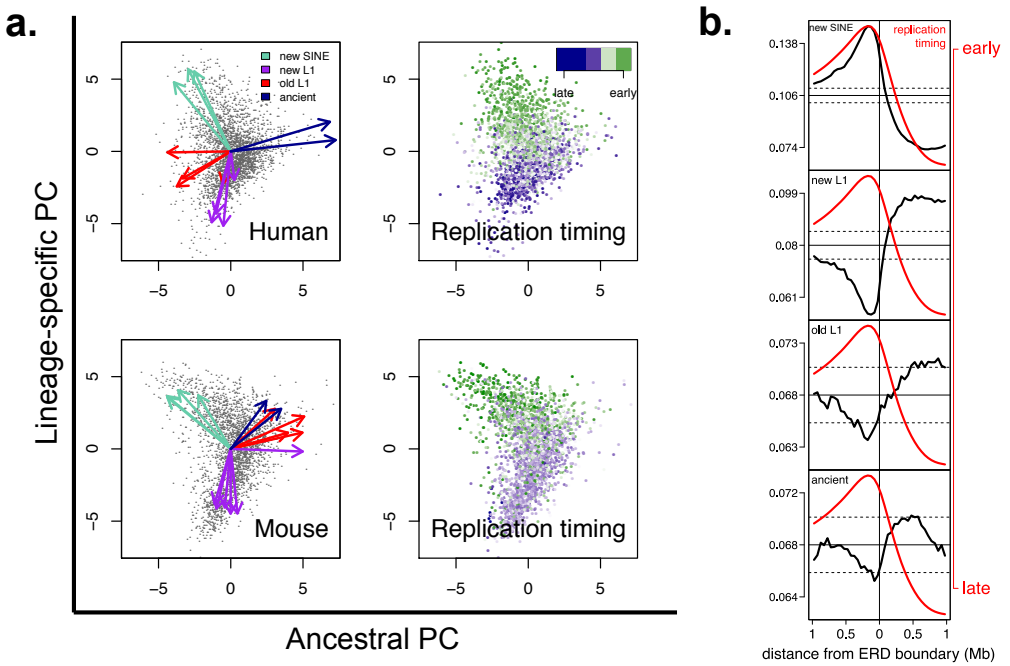


Figure 3. Genomic disttibutions of retrotransposons associate with distinct genomic environments. **a**, PCA of human and mouse retrotransposon content and mean genome replication timing in human HUVEC cells and mouse EpiSC-5 cells. **b**, Retrotransposon density per non-overlapping 50 kb intervals from a pooled set of ERD boundaries across all 16 human cell lines. Black dashed lines indicate 2 standard deviations from the mean (solid horizontal black line). Red line indicates mean replication timing across all samples.

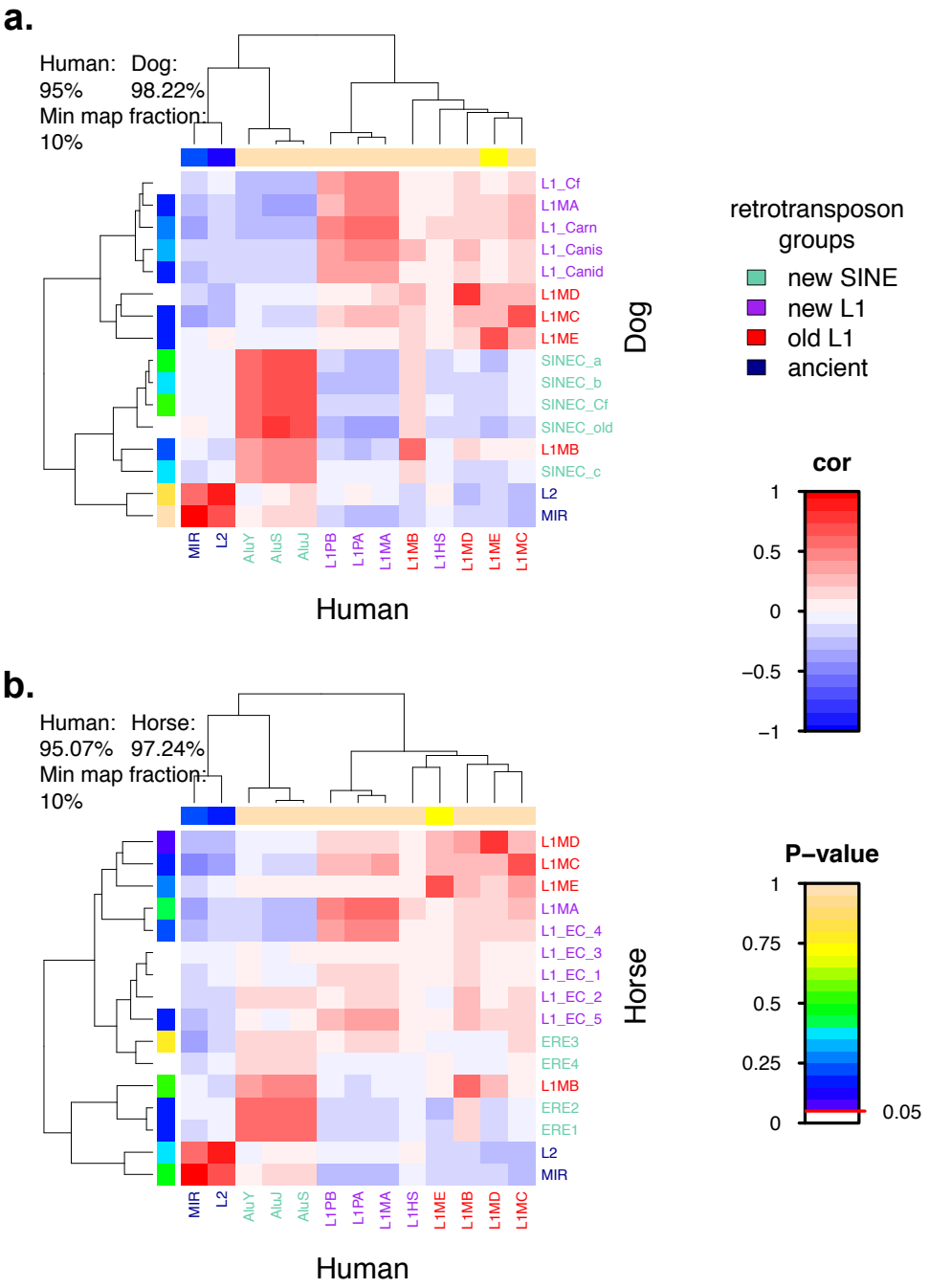


Figure 4. Genome-wide spatial correlations of humanised retrotransposon families. Heatmap colours represent Pearson's correlation coefficient for genomic distributions between humanised **a,** dog and human retrotransposon families, and humanised **b,** horse and human retrotransposon families. Values at the top left of each heatmap reflect the proportion of each genome analysed after filtering at a 10% minimum mapping fraction threshold (Fig. 1a). Dog and horse P-values represent the effect of humanising on filtered non-human retrotransposon density distributions (Fig. 1e). Human P-values represent the effect of filtering on the human retrotransposon density distributions (Fig. 1f).

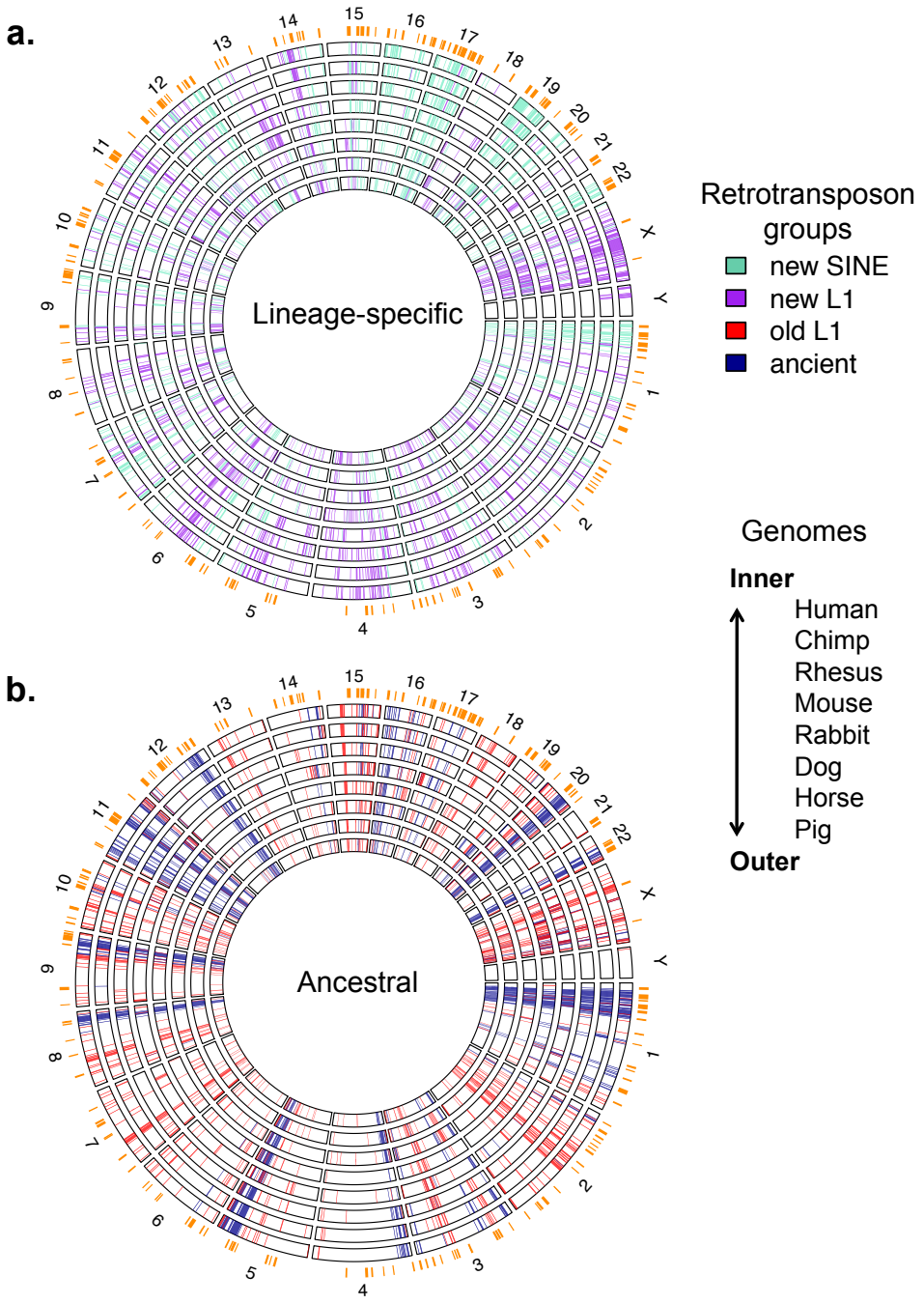


Figure 5. Retrotransposon accumulation patterns are conserved across mammals. a, Top 10% of genome segments based on retrotransposon density of new SINEs and new L1s. **b,** Top 10% of genome segments based on retrotransposon density of ancient elements and old L1s. In both **a** and **b**, segments for non-human genomes were ranked according to their humanised values. Large ERDs (> 2 Mb) from HUVEC cells are marked in orange.

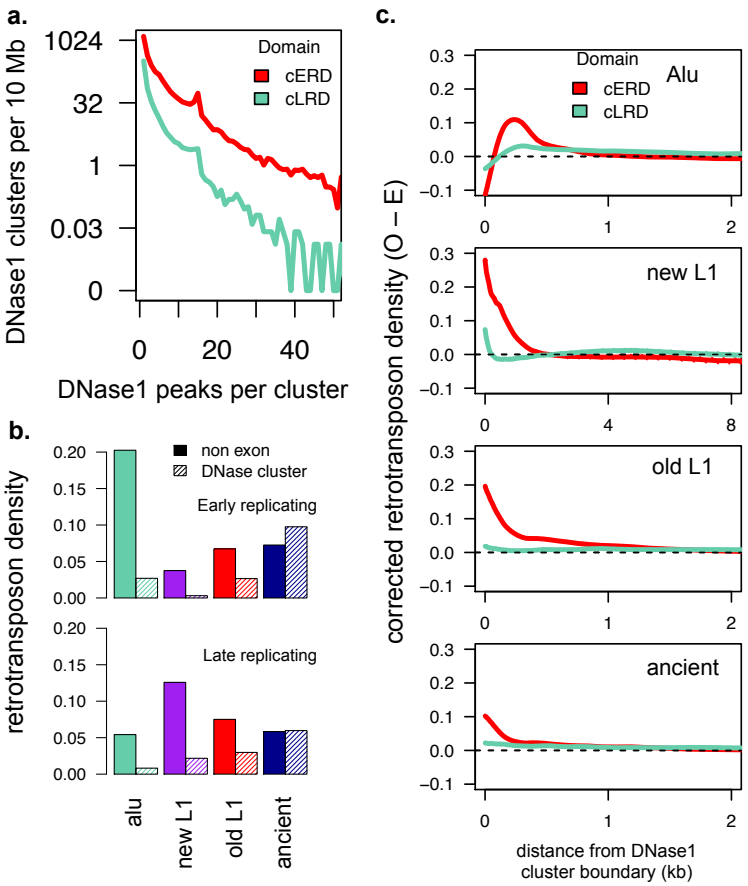


Figure 6. Retrotransposon accumulation occurs in open chromatin near regulatory regions. **a**, The activity of DNase1 clusters in cERDs and cLRDs. DNase1 clusters were identified by merging DNase1 hypersensitive sites across 15 tissues. Their activity levels were measured by the number of DNase1 hypersensitive sites overlapping each DNase1 cluster. **b**, Retrotransposon density of non-exonic regions and DNase1 clusters in cERDs and cLRDs. **c**, Observed minus expected retrotransposon density at the boundary of DNase1 clusters corrected for interval size bias (see methods). Expected retrotransposon density was calculated as each group's non-exonic total retrotransposon density across cERDs and cLRDs. A confidence interval of 3 standard deviations from expected retrotransposon density was also calculated, however the level of variation was negligible.

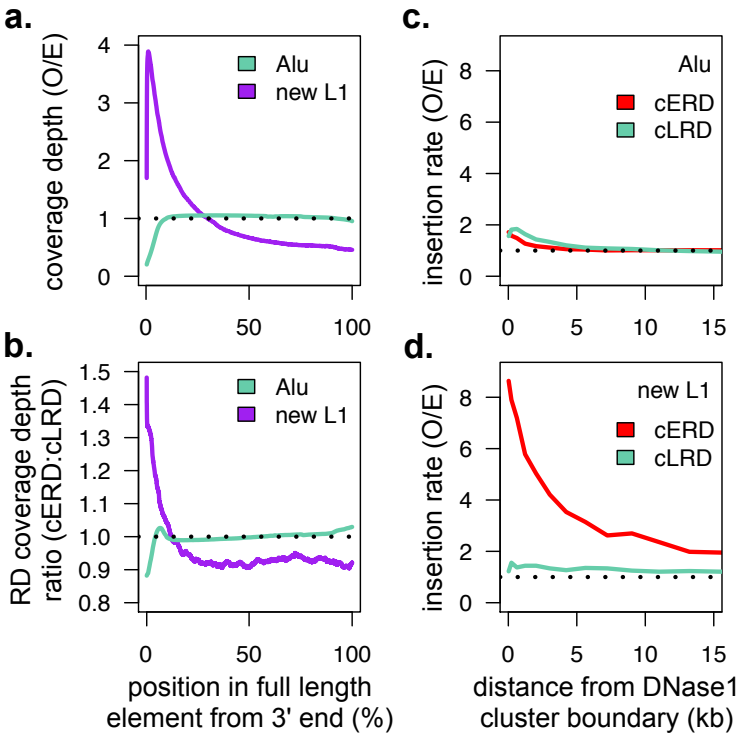


Figure 7. Retrotransposon insertion size is inversely proportional to local regulatory element density. **a**, Observed to expected ratio of retrotransposon position coverage depth measured from consensus 3′ end. Expected retrotransposon position coverage depth was calculated as total retrotransposon coverage over consensus element length. We used 6 kb as the consensus new L1 length and 300 bp as the consensus *Alu* length. **b**, New L1 and *Alu* position density ratio (cERDs:cLRDs). **c**, *Alu* and **d**, new L1 observed over expected retrotransposon insertion rates at DNase1 cluster boundaries in cERDs and cLRDs. Insertion rates were measured by prevalence of 3′ ends and expected levels were calculated as the per Mb insertion rate across cERDs and cLRDs.

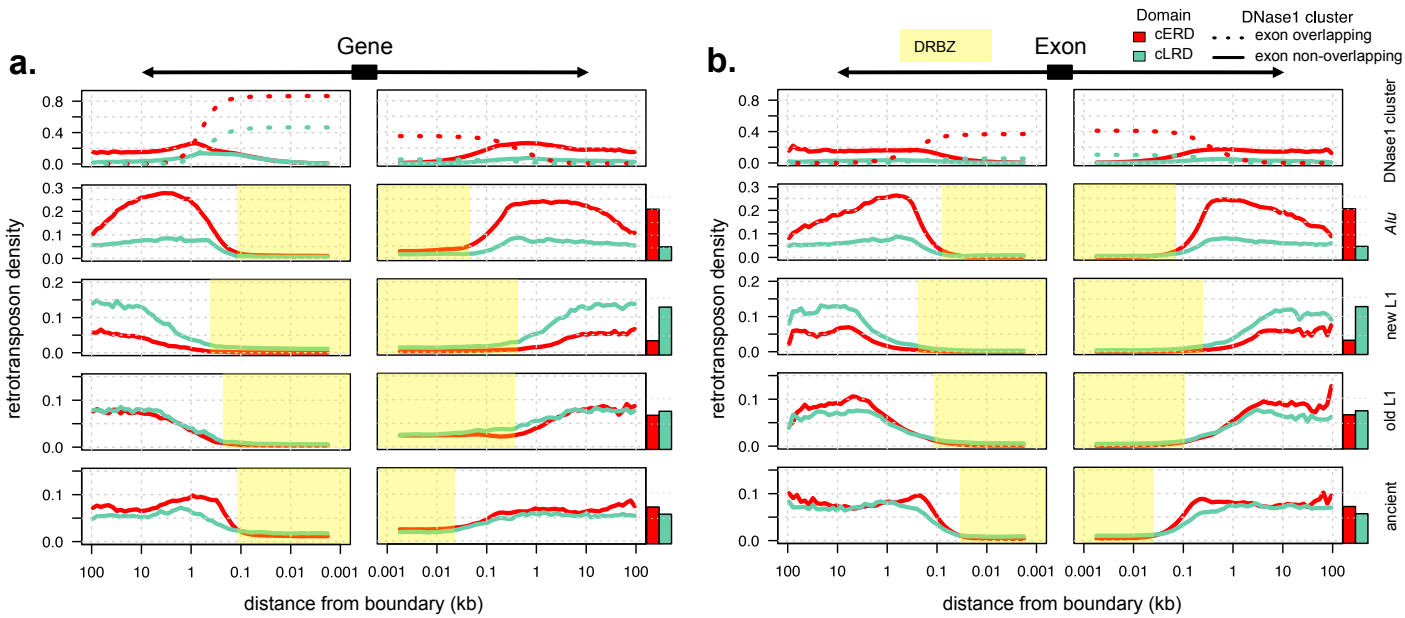


Figure 8. Retrotransposon accumulation within intergenic and intronic regions correlates with the distribution of DNase1 clusters. Density of DNase1 clusters and retrotransposons at each position upstream and downstream of genes and exons in **a**, intergenic and **b**, intronic regions. For DNase1 clusters, dotted lines represent exon overlapping clusters and solid lines represent clusters that do not overlap exons. For retrotransposons, solid lines represent the uncorrected retrotransposon density at exon and gene boundaries. Bar plots show expected retrotransposon density across cERDs and cLRDs. Highlighted regions outline DRBZs, regions extending from the gene or exon boundary to the point where retrotransposon levels begin to increase.

Chapter 3

Divergent genome evolution caused by regional variation in DNA gain and loss in human and mouse

Retrotransposons are responsible for the vast amount of 'new' DNA across mammals, as their accumulation causes genomes to grow in size. However, for genome evolution, DNA gain from retrotransposition is only half the story. Across mammals, DNA loss through deletion occurs at a similar rate as DNA gain through retrotransposon insertion. In this chapter I develop a technique for mapping DNA gain and loss events across distantly related species. I measured regional variation in DNA gain and loss rates across human and mouse and found that different sources of DNA turnover drive lineage-specific evolution of genome architecture. This chapter is in the format of a manuscript that has been submitted to the journal *eLife*.

Statement of Authorship

Title of Paper	Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Reuben M. Buckley, R. Daniel Kortschak and David L. Adelson. Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse. bioRxiv. 2017. Aug 1/179200.

Principal Author

Name of Principal Author (Candidate)	Reuben Buckley
Contribution to the Paper	Processed data, performed analysis, prepared figures and wrote manuscript.
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date: 24/08/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	David L. Adelson
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date: 25/8/2017

Name of Co-Author	R. Daniel Kortschak
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date: 24/8/17

Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse

Reuben M Buckley[1], R Daniel Kortschak[1], David L Adelson[1,*]

1 Department of Genetics and Evolution, The University of Adelaide, North Tce, 5005, Adelaide, Australia

*** david.adelson@adelaide.edu.au**

Keywords: Transposon, Indel, Genome Evolution, Genome Architecture, Human, Mouse

Abstract

The forces driving the accumulation and removal of non-coding DNA and ultimately the evolution of genome size in complex organisms are intimately linked to genome structure and organisation. Our analysis provides a novel method for capturing the regional variation of lineage-specific DNA gain and loss events in their respective genomic contexts. To further understand this connection we used comparative genomics to identify genome-wide individual DNA gain and loss events in the human and mouse genomes. Focusing on the distribution of DNA gains and losses, relationships to important structural features and potential impact on biological processes, we found that in autosomes, DNA gains and losses both followed separate lineage-specific accumulation patterns. However, in both species chromosome X was particularly enriched for DNA gain, consistent with its high L1 retrotransposon content required for X inactivation. We found that DNA loss was associated with gene-rich open chromatin regions and DNA gain events with gene-poor closed chromatin regions. Additionally, we found that DNA loss events tended to be smaller than DNA gain events suggesting that they were more tolerated in open chromatin regions. GO term enrichment in human gain hotspots showed terms related to cell cycle/metabolism, human loss hotspots were enriched for terms related to gene silencing, and mouse gain hotspots were enriched for terms related to transcription regulation. Interestingly, mouse loss hotspots were strongly enriched for terms related to developmental processes, suggesting that DNA loss in mouse is associated with phenotypic changes in mouse morphology. This is consistent with a model in which DNA gain and loss results in turnover or "churning" of regulatory regions that are then subjected to selection, resulting in the differences we now observe, both genomic and phenotypic/morphological.

Introduction

Evolution as a result of natural selection has led to many streamlined forms which follow directly from their function. However, in the case of genome evolution of complex organisms this connection is not quite so direct. One example is the evolution of genome size. In vertebrates, gene content has remained relatively constant, while the fraction of non-coding DNA varies drastically (Gregory 2005; Elliott and Gregory 2015; Gregory 2001). This observation is at the heart of the C-value enigma and raises many questions regarding the molecular drivers and evolutionary impacts of genome size variation. The major factor contributing to the total non-coding DNA genomic fraction is transposon load, due to mobile DNA elements that have actively replicated throughout evolution (Gregory 2001; Elliott and Gregory 2015). In humans, since their divergence from the common placental ancestor, transposon activity has caused approximately 815 Mb of DNA gain, almost one third of their extant genome (Kapusta et al. 2017; Lander et al. 2001). However, this is not the only factor driving genome size evolution. DNA loss via deletion also plays a role, with approximately 650 Mb of the human genome being lost over the same time period (Kapusta et al. 2017). Across mammals and birds these two forces operate in opposition to each other leading to the accordion model of genome evolution, where departures from this DNA gain and loss equilibrium cause genomes to either grow or shrink (Kapusta et al. 2017). Importantly, our understanding of DNA gain and loss stems from genome-wide estimates rather than detection of individual events. Therefore, the role of genome structure on widespread DNA gain and loss and its subsequent impact on lineage-specific species evolution remains unknown.

The 'accordion' model of genome size evolution raises important questions regarding the roles of natural selection and genetic drift. Genome size, like any other heritable trait, is shaped by a combination of both of these factors (Lynch and Walsh 2007). However, the contribution of each mechanism in diverse taxa remains an open question in biology, with evidence to support the impact of each (Whitney and Garland Jr 2010). For genome evolution driven by selection there are observations of various phenotypic correlates consistent across both mammals and birds. One example is the evolution of powered flight in bats and birds which requires a high metabolic rate. Because metabolism is more efficient in smaller cells, it has been suggested that in flying species there is particularly strong selection pressure against genome growth (Wright et al. 2014; Vinogradov and Anatskaya 2006; Kapusta et al.

2017). Alternatively, observed genome size variation can result from neutral evolutionary processes. Many higher order vertebrates have low effective population sizes resulting from reduced efficiency of selection (Lynch and Conery 2003), suggesting that neutral or mildly deleterious mutations such as some transposon insertions can easily reach fixation. Moreover, as transposons quickly accumulate the probability of deletions through non-allelic homologous recombination also increases, counteracting their initial impact on genome growth (Hedges and Deininger 2007; Petrov et al. 2003). Within this context, the accordion model is an emergent property based on transposon accumulation dynamics. Importantly, the signatures of selection for an optimal genome size are not always consistent; the Chinese tree shrew has a high metabolic rate but a relatively large genome of 2.86 GB (Fan et al. 2013). This suggests that the role selection plays in driving genome size evolution is likely taxon-specific. Further, neither mechanism takes into account the underlying genome structure.

The genomic DNA of complex organisms is wrapped around nucleosomes and packaged into various conformations that regulate the access of different gene regulatory factors to their target sites. This hierarchical genome structure means that the impact and likelihood of particular mutations is highly context-specific, resulting in regional variation in both the susceptibility and tolerance to mutations. Here, susceptibility is the likelihood of a mutation occurring and tolerance is the degree to which the mutation does not adversely impact fitness. The observed accumulation patterns of DNA gain and loss events arise from the interaction of region-specific susceptibility and tolerance. For example, small ($\leq$ 30 bp) insertion or deletion (indel) events in the human genome are correlated with recombination rate and are enriched for topoisomerase cleavage sites (Kvikstad et al. 2009, 2007). This suggests that the biological role of certain regions may cause them to be particularly susceptible to indel mutations. In the case of larger events such as transposon insertions, the prevailing model suggests that long interspersed elements (LINEs) accumulate in gene-poor regions where they are most tolerated (Gasior et al. 2007). The evolution of genome size via DNA gain and loss is not only shaped by higher order factors such as cell size and metabolic rate, but is intimately linked to the underlying genome structure.

To better characterise the molecular drivers and evolutionary impacts of DNA gain and loss, we calculated lineage-specific gain and loss rates across the human and mouse genomes. Human and mouse were chosen specifically for three reasons. Firstly, both species have well characterised genomes with highly accurate and well annotated assemblies (Lander et al.

2001; Chinwalla et al. 2002) and have both been used frequently in comparative genomic analyses resulting in many easily accessible pairwise alignment datasets available on the UCSC genome browser (Tyner et al. 2016). This makes it possible to compare them to a wide variety of outgroup species and detect genomic features that associate with DNA gain and loss. Secondly, the mouse genome is significantly smaller than the human genome, making it possible to detect a large number of lineage-specific deletion events (Chinwalla et al. 2002; Laurie et al. 2012). Finally, human and mouse genomes contain similar lineage-specific transposon families (Chinwalla et al. 2002). This means that both species share similar mechanisms for DNA gain, making it easier to compare differences between associations with other types genomic features.

For our analysis, we detected DNA gain and loss events using two distinct, yet complementary, methods from which we characterised DNA gain and loss hotspots. From this we compared the genomic distributions of our hotspots to the genomic distribution of various features associated with genome evolution and genes that participate in particular biological processes. Our results revealed that DNA gains and losses occur in different regions across autosomes, while DNA gains from both species are particularly enriched on the X chromosome where they overlap. DNA gain events generally associate with L1 accumulation and DNA loss occurs in regions associated with biological activity such as transcription and regulation. Although DNA gain and loss in human occurred mostly in different regions, they both tended to impact on the same biological processes, while in mouse DNA loss was enriched for developmental genes and DNA gain did not associate with any particular biological process.

Materials and methods

Net data structure and feature extraction

For feature extraction, nets were obtained from the UCSC genome browser (Kent et al. 2002, 2003). Nets are a common format for representing pairwise genome alignments. Each net contains chained blocks of aligning sequence shared between a reference and a query genome. In order for alignment blocks to be chained together their ordering must be consistent between both genomes. Often gaps between chained blocks can contain smaller chains. It is this hierarchical structuring of the highest scoring chains at the top level with lower scoring

chains filling in alignment gaps that makes nets. Importantly, in the reference genome nets provide only a single layer of coverage. However, two separate nets may occasionally overlap in the query; this is usually caused by segmental duplication in the reference. These conflicts were resolved by discarding all reference nets that did not overlap nets generated from a query reference alignment. Following this filtering process, only reciprocal best hit (RBH) nets remained. In our analysis we referred to alignment blocks within a chain as 'chain-blocks' and the spaces between chain-blocks also within a chain as 'chain-gaps'. The start and end coordinates in both the reference and query genome were recorded for each chain-block and chain-gap. The programs get_gaps_net.go and get_fills_net.go were used to extract all chain-gaps. Regions of chain-gaps that were overlapped by chain-blocks in lower ranked chains were discarded. Additionally, regions that were discarded as non-RBHs or fell outside of nets were plotted against synteny blocks to determine the loci hidden from our analysis in both species. Synteny data was obtained from the synteny portal (`http://bioinfo.konkuk.ac.kr/synteny_portal/`) (Lee et al. 2016).

Identifying ancestral elements

Chain-blocks were extracted from all genomes identified as outgroups to human and mouse. They were combined into a single file and merged using the bedtools genomecov function with the '-bg' option. This process returned a set of potential 'ancestral elements' along with their corresponding coverage depth. To identify false-positives and estimate the type 1 error rate, we used the genomic positions of a set of known lineage-specific repeat families in human and mouse, since lineage-specific repeat insertions should not overlap ancestral elements. The percentage overlap of our lineage-specific repeats set with ancestral elements was measured at each minimum coverage level. A similar approach was used to estimate the type 2 error rate; the type 2 error rate was estimated as the percentage of chain-blocks that did not overlap ancestral elements. To minimise our type 1 errors we selected a minimum coverage depth threshold independently for both hg19 and mm10, where nucleotide positions with coverage depth below the threshold were not considered as ancestral elements. The basis for this approach was that nucleotide positions in our reference genomes that aligned to a large number of outgroup species were highly likely to share ancestry with those species. In contrast, nucleotide positions in our reference genomes that aligned to very few outgroup species were likely errors caused by spurious alignments between complex regions that are

difficult to map. Importantly, reductions in our type 1 error caused an increase in our type 2 error. Therefore, we chose the highest possible minimum coverage threshold, where the gain in the cumulative proportion of type 1 errors from lower threshold values was greater than the gain in proportional increase of type 2 errors.

Identifying recent transposon insertions

For both hg19 and mm10, genomic coordinates for transposons were obtained from the Repeat Masker database (Smit et al. 2015). Based on their overlap with chain-blocks or ancestral elements, individual transposons were classified as either recent or ancestral. In addition to this, the percent divergence from consensus family sequence and the proportion of total sequences of transposon family members that overlapped ancestral elements or chain-blocks were calculated. This data was then used in linear discriminant analysis to build a transposon family classifier. Our classifier was trained using the original individual transposon classifications. After training, entire families were classified as either recent or ancient using the family-wise means of the feature values. Finally, transposons from families classified as recent but overlapping gaps between reference and query were classed as lineage-specific insertions.

Gap annotation and placement

Chain-gaps extracted from nets were annotated as either DNA gain or DNA loss based on two distinct yet complementary annotation methods; the recent transposon-based method and the ancestral elements based method. The ancestral element-based method infers the ancestral state of a gap. For example, an mm10 gap overlapping an ancestral element would be annotated as an mm10 loss, whereas the same gap not overlapping an ancestral element would be annotated as an hg19 gain. The recent transposon-based method instead identifies DNA gains. In this case an mm10 gap overlapping a recent transposon would be annotated as an hg19 gain, while an mm10 gap not overlapping a recent transposon would be annotated as an mm10 loss.

After all chain-gaps between a reference and query were annotated in both genomes, the remaining non-aligning sequences were ‘placed’ in the genomes they were absent from. This process is referred to as ‘gap placement’ and is performed on the non-aligning sequence of chain-gaps that remain in the reference genome after a reference query alignment. These

non-aligning reference sequences are absent from the query and are either the result of DNA gain in the reference or DNA loss in the query. Using the coordinate mappings of the $5'$ and $3'$ adjacent chain-blocks of each chain-gap, the non-aligning reference sequence of a chain-gap is inserted into the query genome at the corresponding position, where placed gaps are oriented relative to the genome they are placed in. Importantly, gap placement begins by placing chain-gaps at the bottom chain level of nets and ends by placing chain-gaps at the top chain level. This process ensures that non-aligning sequence in overlapping chain-gap annotations caused by hierarchical structure of nets are only placed once. Once the corresponding position of a gap has been identified, the downstream query coordinates are incremented by the size of the annotated chain-gap being placed. This creates a synthetic genome consisting of DNA gains and losses that occurred across both the reference and query lineages. The total length of our synthetic genomes is equal to the total length of the query genome and the total length of annotated chain-gaps from the reference. Finally, the synthetic genomes were segmented at a window size of 200kb into distinct genomic bins where the total size of each gap annotation was tallied. Genomic bins with less than 150 kb that did not belong to assembly gaps or non-RBH regions were discarded. Importantly, our decision to use a synthetic genome meant that placed chain-gaps larger than our window size would spread across window boundaries, ensuring that genomic bins would contain no more than 200 kb of sequence.

Hotspot identification

Hotspots for reference gain, reference loss, query gain and query loss in both hg19 and mm10 were identified using the Getis-Ord local statistic found in the R package 'spdep' (Bivand et al. 2013; Bivand and Piras 2015). The Getis-Ord local statistic for genomic bin i is calculated as:

$$G_i^* = \frac{\sum w_{i,j}x_j - \bar{X}\sum w_{i,j}}{S\sqrt{\frac{n\sum w_{i,j}^2 - \left(\sum w_{i,j}\right)^2}{n-1}}}, \quad (1)$$

where x_j is the number of bp belonging to a particular gap annotation within bin j, $w_{i,j}$ is the spatial weight between bin i and j, n is the number of bins for a particular genome, $\bar{X} = \frac{\sum x_j}{n}$ and $S = \sqrt{\frac{\sum x_j^2}{n} - \bar{X}^2}$ (Getis and Ord 1996). For the neighbourhood weight matrix W, $w_{i,j}$ was given a spatial weight of 1 if bin i and bin j were considered neighbours. For bin i and j to be considered neighbours bin j had to be within 600 kb of bin i. After

calculating G_i^* for each bin and each gap annotation in both genomes, all G_i^* values were converted to P-values and adjusted for multiple testing using the false discovery rate (FDR). Bins were only considered hotspots if their G_i^* was > 0 and had a FDR < 0.05.

Obtaining genomic features

A set of genomic features was obtained from a range of sources to identify factors potentially driving DNA gain and loss. GC content was calculated as the proportion of chain-blocks per bin using the hg19 and mm10 Biostrings-based genome R packages (Team TBD 2014a,b; Pages 2017). CpG islands for both hg19 and mm10 were obtained from the UCSC genome browser (Tyner et al. 2016). DNaseI hypersensitivity (DNaseI HS) peaks for hg19 were obtained from UCSC as part of the DNaseI master track (`http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseMasterSites/`). The master track was generated by combining DNaseI HS sites from across 125 cell lines produced by the University of Washington and Duke University ENCODE groups (ENCODE Project Consortium et al. 2012). The Individual cell line data can be located using the accessions GSE29692 and GSE32970. DNaseI HS peaks for mm10 were obtained from UCSC as individual samples mapped to mm9 (`https://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9\&g=wgEncodeUwDgf`). Individual peaks from each sample were merged into a single file, creating a single set of DNase1 HS peaks. The merged mm9 peaks were then converted to the mm10 assembly using the UCSC liftover tool (Hinrichs et al. 2006). Mouse DNaseI HS peaks were generated using DNaseI digital genomic foot-printing performed by the University of Washington ENCODE group (ENCODE Project Consortium et al. 2012). This data can also be obtained using the accession GSE40869. Recombination rates for human were identified as part of the HapMap project (`ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/`)(International HapMap Consortium et al. 2007). However, recombination hotspots were only available for earlier phases of the HapMap project (`ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2006-10_rel21_phaseI+II/hotspots/`). The hotspots were initially mapped to hg17 and then converted to hg19 coordinates using the UCSC liftover tool. Recombination hotspots were identified using the methods outlined in Winckler et al. (2005) and McVean et al. (2004). Recombination rates and hotspots in mouse were calculated in mm9 based on two separate datasets (Brunschwig et al. 2012; Kirby et al. 2010; Yang et al. 2011). They were converted to mm10 using the UCSC liftover

tool. Importantly, recombination data was only available for mouse autosomes. During enrichment tests this was taken into account by removing the sex chromosomes from the sample space. Exons and introns for both hg19 and mm10 were extracted from UCSC genome annotations available from TXDB R packages (Carlson 2015, 2016; Lawrence et al. 2013). Retrotransposon coordinates for hg19 and mm10 were obtained from the Repeat Masker database (`http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html`) (Smit et al. 2015). The Repeat Masker version used for hg19 and mm10 was open-4.0.5 with repeat library 20140131. Retrotransposons were sorted into the following categories: ancient elements, ancestral L1s, lineage-specific L1s and lineage-specific SINEs using prefixes for families of known lineage-specific and ancestral activity (Giordano et al. 2007). Ancient elements were identified by the class names 'SINE/MIR' and 'LINE/L2'. Ancestral L1s were identified using the family name prefixes 'L1ME', 'L1MD', 'L1MC', 'L1MB' and 'L1MA'. Human lineage-specific L1s were identified using the family name prefixes 'L1PB', 'L1PA' and 'L1HS'. Mouse lineage-specific L1s were identified using the family name prefixes 'Lx', 'L1Md', 'L1_Mus', 'L1_Mur' and 'L1_Mm'. Human lineage-specific SINEs were identified using the family name prefix 'Alu'. Mouse lineage-specific SINEs were identified using the family name prefixes 'PB', 'B1', 'B2', 'B3' and 'B4'. Lamina associated domains (LADs) for hg19 were obtained from the UCSC genome browser (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/laminB1Lads.txt.gz`) (Guelen et al. 2008). LADs for mouse were constitutive across several samples and were obtained using the accession GSE17051, they were converted from mm9 assembly to mm10 assembly using the UCSC liftover tool (Peric-Hupkes et al. 2010). For each feature, except recombination rate, the per 200 kb coverage level for each bin was calculated. For recombination rate the mean rate per bin was used.

Genomic feature enrichment

Feature enrichment was detected on the basis of a permutation test. For each feature and hotspot in both hg19 and mm10, a background distribution was generated by calculating the difference in means between a set of resampled hotspot and non-hotspot bins 10,000 times, resampling was performed without replacement. The background distribution was then used to convert the differences in means between observed hotspot and non-hotspot bins into a Z-score to allow standardisation between features and gap annotations and provide the

direction of the association.

GO term enrichment analysis

Gene ontology (GO) term enrichment was calculated using the topGO package in R (Alexa and Rahnenfuhrer 2016). Genes within each hotspot region were independently tested against the genomic background. For enrichment, the Fisher test was used in combination with four separate algorithms: the classic algorithm treats each term independently whereas the elim, weight and parent-child algorithms factor in the GO inheritance structure (Alexa et al. 2006; Grossmann et al. 2007; Ashburner et al. 2000); the elim algorithm removes all genes annotated to a significantly enriched GO term from all of the terms ancestors; the weight algorithm behaves similarly, instead of removing genes from the ancestors of enriched GO terms, it creates a more subtle effect by reducing the weight of genes annotated to the ancestors of enriched GO terms (Alexa et al. 2006); for the parent-child algorithm, the enrichment score for a particular term takes into account the probability a random set of genes of the same size contains the same exact parents (Grossmann et al. 2007). Because these algorithms adjust the enrichment probabilities they obviate the need to account for multiple testing (Alexa and Rahnenfuhrer 2016).

Software and data analysis

All statistical analyses were performed using R including the packages GenomicRanges, RMySQL, dplyr and Bioconductor (R Core Team 2015; Lawrence et al. 2013; Ooms et al. 2016; Wickham and Francois 2015; Gentleman et al. 2004). Code used to perform analyses can be found at: `https://github.com/AdelaideBioinfo/regionalGenomeTurnover`.

Results

Detecting DNA gain and loss events.

Across genomes and throughout evolution DNA is frequently gained and lost by the processes of insertion and deletion. To identify individual events and quantify DNA gain and loss at a regional level in hg19 and mm10, we obtained pairwise alignment data between both genomes in the form of nets from the UCSC genome browser (methods) (Tyner et al. 2016;

Kent et al. 2003). By taking advantage of the data's hierarchical structure we were able to estimate DNA gain and loss in regions that have undergone rearrangements. We processed our data in three distinct steps; 1) extract features (Fig. 1a), 2) annotate gaps (Fig. 1b-c) and 3) place gaps (Fig. 1d).

For step 1, chain-gaps and chain-blocks were extracted from nets considering only chain-gaps of at least 10 bp in size (Fig. 1a) (methods). Our approach allowed us to keep track of each feature's position in both the reference and query genome. This is especially important since it is not possible to identify deletions when the corresponding coordinates between species are lost. After extracting features we found that approximately 111 Mb of hg19 and 174 Mb of mm10 were not contained within nets (Table 1). Alignment gaps that didn't belong to any nets in human and mouse tended to overlap regions between two conserved synteny blocks (Fig. S1-S2). With the remaining features extracted from hg19 and mm10, we used the corresponding coordinates between reference and query to identify features that were reciprocal best hits (RBHs). This removed features in the reference genome that mapped to similar locations in the query, which are likely the result of segmental duplication. After filtering out non-net and non-RBH regions, 1014.3 Mb of chain-blocks and 1465.8 Mb of chain-gaps remained in hg19, and 994.4 Mb of chain-blocks and 1191.5 Mb of chain-gaps remained in mm10 (Table 1). Since our processed nets for each genome are supposed to only contain RBH features, it is expected that the coverage of chain-blocks should be equal between hg19 and mm10. To determine the source of this discrepancy, we analysed the number of chain-gaps bellow our minimum size cut off and found that when these were taken into consideration the difference in chain-block size was reduced to approximately 1 Mb.

Next, for step 2 we annotated chain-gaps as either lineage-specific DNA gain or DNA loss. To annotate gaps we used two complementary methods, an ancestral elements-based method and a recent transposon-based method. The ancestral element-based method uses outgroup species to annotate gaps by inferring their ancestral state (Fig. 1b). For example, if a particular sequence between a reference and outgroup is conserved but presents as a gap in the query it is likely that this sequence was lost from the query. Alternatively, if this particular sequence in the reference presents as a gap in both the query and the outgroup it is likely that this sequence was instead gained in the reference. An important consideration for identifying ancestral elements is the type 1 (false positive) and type 2 (false negative) error rates, where type 1 errors are lineage-specific regions annotated as ancestral elements and

type 2 errors are ancestral regions annotated as lineage-specific. To reduce our type 2 error rate we obtained the genomes of a large range of human and mouse outgroup species from the UCSC genome browser (Table S2). Across all of our outgroup species we extracted all the chain-blocks and merged overlapping intervals to create our ancestral elements. This strategy increased the chance of finding ancestral DNA in our reference that may have been lost in one or more of our outgroup species. For both hg19 and mm10 we found that total genome coverage of ancestral elements reached asymptotic levels at approximately 18 outgroup species (Fig. S3). However, this strategy also came with the trade-off of increasing our type 1 error rate. To control error rates we measured how type 1 and type 2 errors responded to changes in coverage depth of outgroup chain-blocks at each position in hg19 and mm10 (Fig. S4). Based on these results we annotated human ancestral elements at an outgroup coverage depth ≥ 6 and mouse ancestral elements at an outgroup coverage depth ≥ 4 (Fig. S4). This strategy removed $> 85\%$ ancestral elements overlapping known lineage-specific repeats in mouse and $> 95\%$ of ancestral elements overlapping known lineage-specific repeats in human. For remaining chain-blocks, we found that 94.2% in human and 85.2% in mouse were supported by our annotated ancestral elements (Table 1). Our very low error rate in human indicates that we were able to accurately determine the amount of mm10 DNA loss and hg19 DNA gain. However, our error rates in mm10 suggest that ancestral regions alone are insufficient to accurately estimate hg19 DNA loss and mm10 DNA gain.

To complement and overcome potential shortcomings of the ancestral element-based method of estimating DNA gain and loss, we adopted a recent transposon-based method. We identified transposon families with lineage-specific activity and used them to annotate gaps as lineage-specific DNA gain or loss (Fig. 1c). For example, recent transposon sequences in hg19 that overlap gaps in mm10 are annotated as hg19 gains, where ancestral transposon sequences in hg19 that overlap gaps in mm10 are annotated as mm10 losses. This approach has been used previously to identify DNA loss in the mouse and human lineages (Chinwalla et al. 2002; Hardison et al. 2003).

In order to annotate gaps using the recent transposon method, we first had to identify transposon insertions that occurred after mouse and human diverged from their common ancestor. Because transposon families have undergone distinct bursts of activity at particular points in time, we decided to classify transposon families as either 'recent transposons' or 'ancestral transposons', and use members of those respective classifications to annotate

our chain-gaps. The main challenge in this approach is identifying lineage-specific activity of transposons. Generally, transposon families are considered to be ancestral transposon families when they are shared between two species. However, there is a possibility some ancestral transposon families may have been active during the period of human and mouse divergence and continued replicating in each lineage independently. This means families that would have been otherwise classified as ancestral transposons may have actually undergone varying amounts of lineage-specific transposition.

To overcome the problem of misclassifying the activity of otherwise ancestral transposon families, we used linear discriminant analysis to build a transposon family classifier for both human and mouse. We initially obtained transposon coordinates from the Repeat Masker database and classified individual transposons as 'ancestral transposons' if they overlapped ancestral elements or chain-blocks and as 'recent transposons' if they did not. Next, we trained our classifier using two separate variables. The first variable was each transposon's percent divergence from their family consensus sequence, often used as an indicator of transposon age (Kapitonov and Jurkal 1996; Smit et al. 1995). The second variable was the proportional overlap between each transposon family and ancestral elements or chain-blocks as measured by bp coverage. After training we used our classifier to group each family based on the family-wise means for the variables above (Fig. S5). We identified 656 recent human transposon families and 689 recent mouse transposon families. Our results suggest that at least 176 families were active during human and mouse divergence leading to a mixture of both ancestral and lineage-specific insertions (Table S1). Moreover, the percent divergence of these families is consistent with transposon activity occurring after the evolution of ancestral transposons and prior to the evolution of lineage-specific transposons (Fig. S6). Surprisingly, we also identified some transposon families that were not shared between human and mouse, and yet were annotated as ancestral. However, these families were usually small and together they covered less than 1 Mb of their respective genomes (Table S1). In addition, our results for mm10 indicate potential drawbacks in using the ancestral element-based method for annotating gaps; percent divergence from consensus for some recent transposon families is similar to ancestral transposon families. While this is consistent with an elevated rate of substitution in the rodent lineage, it suggests that a large number of regions in mm10 that share ancestry with our outgroup species may have diverged beyond the alignment threshold (Fig. S5). Collectively, these results demonstrate

the difficulty of identifying recent transposon insertions based on family name alone. For this reason we decided to annotate chain-gaps using our newly classified recent transposon families, which were classified using a combination of family-wide and transposon-specific factors in conjunction with comparative genomic approaches.

Using both the ancestral element and recent transposon based methods, we annotated a large number of chain-gaps with varying levels of consistency. In hg19, both methods were largely consistent in identifying human-specific DNA gains and mouse-specific DNA loss. However, in mm10 there was less agreement between the methods; while the majority of mouse lineage-specific DNA gains identified by both methods tended to overlap, the majority of human lineage-specific DNA loss did not (Table 2). This is mostly likely due to limitations for detecting ancestral elements in mm10. We found that only 85% of mm10 chain-blocks were supported by ancestral elements as opposed to 95% in hg19 (Table 1), suggesting that many ancestral elements were not identified using our outgroup species. This is a key weakness in our approach; if there is an underlying error for detecting human DNA loss in mm10, it means that we would also be overestimating the amount DNA gain in mm10. However, by using two distinct yet complementary methods, we are able to identify potential sources of error and estimate their impact. One explanation for missing ancestral elements may be that DNA gain and loss events that occurred in either the mouse or human clade overlap DNA gain and loss events that occurred across a large number of our outgroup species. However, as stated above, nucleotide divergence rates may also play a role. Some regions in mm10 may have diverged so much that it is impossible to perform a pairwise alignment with our outgroup species. Despite the above mentioned inconsistencies between the methods in mm10, it is clear that the amount of DNA loss in human is much smaller than the amount of DNA loss in mouse and the amount of DNA gain for both. The difference in loss rates for human and mouse is mostly consistent with a high deletion rate in the mouse genome that has caused it to shrink in size since divergence with human (Chinwalla et al. 2002; Laurie et al. 2012).

To further characterise the results from each method we compared the length distributions of their gap annotations. For DNA gain events in hg19 and mm10, the ancestral element method displayed a much higher frequency of small elements than the recent transposon method. This may be caused by spurious alignments between similarly structured recent transposons found in reference and outgroup species, effectively separating the annotation

gain events into smaller pieces. Moreover, the recent transposon method identified much higher frequencies of DNA gain events that correspond to full length consensus sequences of known transposon families (Fig. 2a-2b). Conversely, the length distributions for DNA loss events identified by each method were much more similar, especially in mm10. In hg19 the frequency of events detected by the ancestral element method were much lower than those detected by the recent transposon method (Fig. 2c-2d). This is consistent with the low number of ancestral elements in the mouse genome. However, the high level of consistency for both methods in identifying hg19 DNA gain and mm10 DNA loss where there is good support for outgroup species is highly encouraging. It indicates that the recent transposon method is a reasonably effective method in identifying DNA gain and loss in species where it is difficult to detect ancestral elements. Consistent between both methods is size distribution difference between DNA gain and loss. DNA gain events are mostly over 100 bp in length while DNA loss events are mostly under 100 bp.

In both hg19 and mm10 we annotated a large number of gain and loss events using two distinct methods. However, to measure the total amount of DNA turnover at particular loci, gaps annotated in a query genome needed to be mapped to a reference genome. Hence, gap annotations were placed using the reference and query coordinates we extracted from our nets in step 1 (methods) (Fig. 1d). To account for the placement of gaps from one genome into another, we adjusted the genomic coordinates at the target loci, resulting in a synthetic genome for both species (methods). Each synthetic genome contains both hg19 and mm10 annotated gaps in either an hg19 or mm10 genomic background. Finally, our resulting dataset consists of 4 synthetic genomes; mm10 with gap annotations based on the ancestral element method, mm10 with gap annotations based on the recent transposon method, hg19 with gap annotations based on the ancestral element method and hg19 with gap annotations based on the recent transposon method. Collectively, these results demonstrate that it is possible to identify locations for the majority of DNA gain and loss events since human and mouse divergence. Using our identified DNA gain and loss events it is possible to characterise genome-wide patterns of DNA gain and loss and to begin to determine how DNA turnover may impact on mammalian genome evolution.

Genome-wide characteristics of DNA gain and loss.

Genome size evolution in mammals follows an accordion model, where DNA gain is counteracted by DNA loss to maintain a relatively constant genome size (Kapusta et al. 2017). To characterise how DNA gain and loss interacts with genome structure, we used our synthetic genomes to analyse the genomic distribution of DNA gain and loss events in hg19 and mm10. We began by segmenting synthetic genomes into 200 kb non-overlapping bins and tallying the total bp coverage of each type of gap annotation. Bins with less than 150 kb of DNA not belonging to RBH nets were removed and our tallies were normalised to reflect DNA gain and loss amounts per 200 kb. Because gap annotations from both species can be placed within a single genome, we are able to directly compare their genomic distributions.

Using our binned synthetic genomes we compared the variation and average amount of regional DNA gain and loss identified using each method. Our results showed that variation in regional DNA gain or loss was reasonably consistent across both methods (Fig. 3). For DNA gain this was also quite large, in 200 kb genomic bins the amount of DNA gain in human and mouse spanned a range greater than 70 kb, indicating that some regions underwent much greater levels of DNA gain than others. While bin-wise variation in gain and loss rates was consistent across methods, the average amount of DNA turnover was not. This makes it difficult to reliably calculate the regional amount of DNA turnover or genome growth. However, despite these inconsistencies, bin-wise levels of DNA gain and loss were highly correlated across all cases, with the exception of hg19 DNA loss (Fig. 3a, S7-S8). Following this, we investigated regional DNA gain and loss dynamics by identifying DNA gain and loss genomic hotspots. Hotspots were identified by calculating G_i^* for each bin (methods). We converted our G_i^* values to P-values and calculated the false discovery rate (FDR). Bins whose G_i^* was positive with FDR < 0.05 were considered hotspots. Hotspots were identified for each type of gap annotation found using both gap annotation methods in both synthetic genomes. We found that the size of the hotspot overlap between each gap annotation method for hg19 gain, mm10 gain and mm10 loss was larger than the sum of non-overlapping hotspots (Fig. 3b). Using the hotspot intersect between gap annotation methods, we further characterised regional variation of DNA gain and loss across hg19 and mm10. For the remainder of the analysis the terms 'DNA-gain hotspots' and 'DNA-loss hotspots' refer to the hotspot intersect between each gap annotation method, except for hg19

DNA-loss hotspots which instead refer to hg19 DNA-loss hotspots identified through the recent transposon method. For mm10 DNA loss, mm10 DNA gain and hg19 DNA gain, the intersect was used as it provided a sample of genomic regions where regional DNA gain and loss dynamics were highly supported by both methods. For hg19 DNA loss we used hotspots that were identified using the recent transposon method because the ancestral based method was shown to largely underestimate the total amount of ancestral DNA.

Regional patterns of DNA gain and loss indicate lineage-specific divergence.

The accordion model of genome evolution suggests DNA gain and loss is largely balanced across the entire genome. Whether the individual events are balanced at the local scale remains unknown. We analysed the genomic distribution of hg19 and mm10 gain and loss hotspots by focussing on the within species overlap and the across species overlap. The within species overlap was designed to investigate whether DNA gain and loss is balanced on a regional level, indicating that despite large amounts of DNA turnover, local genome structures stay intact. The across species overlap was designed to investigate whether DNA gain and loss associated with lineage specific divergence in genome architecture. We found that almost 4% of human loss hotspots overlapped human gain hotspots and approximately 6% human gain hotspots overlapped human loss hotspots (Fig. 4,S9). These results showed that DNA gains and losses in human at a regional scale have occurred independently. Conversely, less than 1% of gain and loss hotspots in mouse overlapped each other, with a significant negative association. These results suggest that regional DNA gain and loss in both species is largely unbalanced. For the across species comparison, we found significant levels of overlap between DNA-loss hotspots and negative associations between all other hotspot types at varying levels of statistical significance depending on genomic background. This demonstrates that DNA loss dynamics in both hg19 and mm10 share some degree of conservation while DNA gain dynamics are mostly lineage-specific, suggesting that the acquisition of new DNA may be driving lineage-specific divergence of genome structure.

To further characterise the distribution of hg19 and mm10 gain and loss hotspots, we plotted them against both genomic backgrounds. hg19 and mm10 gain hotspots were most enriched on chromosome X (Fig. 4,S9). This is consistent with chromosome X as a hotspot

for L1 insertion, a particularly large transposon with high levels of lineage specific activity that contributes to X inactivation (Chow et al. 2010). For gain and loss hotspots themselves, hg19 gain hotspot regions were much more dispersed than other types of hotspot region (Fig. 4,S9). Since DNA loss across both species overlaps significantly, this adds to the lineage-specific behaviour of DNA gain dynamics, where regional DNA gain in mouse is more concentrated than in human. Interestingly, DNA loss hotspots in the hg19 genomic background appear more concentrated towards telomeres, suggesting that chromosomal location may play a role in DNA loss dynamics (Fig. 4). However, it is worth noting that this observation did not occur in the mm10 genomic background (Fig. S9). One explanation is that telomeres in mouse are quite recent as mouse chromosomes have undergone a high frequency of breakage and fusion events since divergence from a common ancestor (Murphy et al. 2005). Together, our results demonstrate that regional lineage-specific DNA gain and loss dynamics are relatively context-specific.

Next, we examined whether gain and loss hotspots were correlated with a range of genomic features. The genomic features we analysed are non-randomly distributed and known to play various roles in genome biology. By investigating their association, we may begin to develop insight into the molecular drivers of DNA turnover. To measure the correlation between genomic features and particular gap annotations we performed feature enrichment analysis with 10,000 permutations (methods). The analysis was performed for both mm10 gain and loss and hg19 gain and loss in both the genomic backgrounds. Using both genomic backgrounds we were able to analyse the genomic features from regions in a query genome that have been deleted from a reference. We specifically chose genomic features that could be found in both genomes as indicators for distinct aspects of genome biology. Intron density, exon density, DNaseI hypersensitivity (DNaseI HS) peaks, CpG islands, GC content and lamina-associated domains (LADs) are all indicators of genome activity (ENCODE Project Consortium et al. 2012; Tyner et al. 2016; Guelen et al. 2008; Peric-Hupkes et al. 2010). Most of these features, excluding LADs, are associated with gene dense areas and are linked to their expression or regulation (Thurman et al. 2012). LADs themselves are instead associated with gene-poor regions and gene silencing (Guelen et al. 2008; Peric-Hupkes et al. 2010). We also investigated various groups of transposons whose genomic distributions have been previously characterised and used to investigate genome-wide DNA gain and loss rates. Lineage-specific L1s and SINEs are both major sources of DNA gain via retrotransposition, they both also

have distinct accumulation profiles that are similar across both species (Chinwalla et al. 2002). Lineage-specific L1s tend to accumulate in gene-poor regions while lineage-specific SINEs accumulate in gene rich regions. Ancestral L1s, and ancient elements (MIRs and L2s) have been used previously to indicate levels of DNA loss. Since these elements inserted prior to species divergence, they both provide signatures of ancestral DNA. Differences in the numbers of these elements in similar regions across species can indicate DNA loss (Chinwalla et al. 2002; Laurie et al. 2012). Finally, we investigated the genomic distribution of recombination hotspots and genome-wide profiles of recombination rates (International HapMap Consortium et al. 2007; Brunschwig et al. 2012). We considered recombination as an indicator of genome instability, as meiotic recombination increases the potential for heritable genomic rearrangements (Berg et al. 2010). Importantly, it is worth noting that recombination hotspots and recombination rates in mm10 are autosomal only. This was due to limited data availability for mouse.

Among our features we observed distinct profiles for DNA gain and loss that were largely consistent across both genomes. For DNA loss from both genomes and in both genomic backgrounds we found a strong positive associations with indicators of gene-rich/active genomic regions. This is surprising as biologically active genomic regions are likely to contain many important functional elements. However, it has recently been shown that these regions are particularly prone to genomic instability leading to evolutionary genomic rearrangements (Berthelot et al. 2015). This also suggests the DNA loss is linked to an open chromatin state as it is strongly negatively associated with LADs. In the hg19 genomic background we also found that ancient elements were positively associated with mm10 DNA loss. While ancient elements have been used as indicators of DNA loss, we did not expected they would be quite so strongly associated with it. Moreover, in hg19 ancient elements are negatively associated with DNA loss and have been predicted to play important roles in gene regulation (Kamal et al. 2006). In addition, the high DNA loss rate in these regions may lead to overestimates of the genome-wide DNA loss rate in mouse, as these elements have previously been used as markers for calculating deletion rates (Lander et al. 2001; Chinwalla et al. 2002). Our results also showed that DNA loss in hg19 and mm10 in the hg19 genomic background was positively associated with genomic recombination. This is consistent with previous analyses that have identified an association between DNA loss and recombination (Nam and Ellegren 2012). Interestingly, we did not observe any association with recombination in the mm10 genomic

background. This may be due to the decreased resolution used to calculate recombination rates and identify recombination hotspots in mouse compared to human (Brunschwig et al. 2012; International HapMap Consortium et al. 2007). For DNA gain hotspots we found that their associations with genomic features was less consistent across both species than DNA loss hotspots. For sources of DNA gain, mm10 and hg19 DNA gains were both positively associated with lineage-specific L1s. However, while lineage-specific SINEs were associated with hg19 DNA gain, in mm10 they were associated with DNA loss. This paradoxical finding is likely caused by two separate contributing factors. The first is that lineage-specific SINEs in mouse are not a major contributor to DNA gain compared to human, as their overall coverage levels are much lower (Chinwalla et al. 2002). The second is that lineage-specific SINEs accumulate in gene-rich open chromatin areas which also happen to strongly associate with DNA loss (Buckley et al. 2017). These differences in sources of DNA gain may explain divergence patterns in both species DNA gain dynamics; lineage-specific SINEs are associated with gene-rich/active genomic regions and lineage-specific L1s are associated with gene-poor silent regions such as LADs. Ultimately, this suggests that DNA is accumulating/turned over in different regions at different rates by otherwise conserved mechanisms of DNA gain. Collectively, our results show that DNA gain and loss is associated with specific genomic contexts, leading to differences in genome structure.

DNA gain and loss is non-random and may be a function of mammalian genome structure. However the evolutionary impact of DNA gain and loss is mainly determined by whether or not it affects particular phenotypes. To identify potentially impacted phenotypes we performed gene ontology (GO) enrichment analysis on genes in DNA gain and loss hotspots for biological process GO terms (Ashburner et al. 2000). Because we are interested in identifying whether DNA gain and loss may have driven lineage-specific divergence we compared the significance levels of GO term enrichment between our hotspot types. To do this we performed correlation analysis using the -log10 P-values for GO term enrichment as determined using a Fisher test combined with the 'classic' GO term enrichment algorithm (methods) (Alexa and Rahnenfuhrer 2016). Surprisingly our results showed the highest level of similarity between hg19 DNA gain and hg19 DNA loss (Fig. 6,S10). This is interesting because the overlap between hg19 gain and loss was not statistically significant (Fig. 4, S9). Moreover, when we compare hg19 DNA loss with mm10 DNA loss; gap annotations with a significant degree of overlap (Fig. 4, S9), we found that GO terms were not as similar,

particularly in the mm10 genomic background (Fig. S10). Alternatively, enriched GO terms found in mm10 DNA gain hotspots appeared distinct from GO terms enriched in other DNA gain and loss hotspots. These results echo our above findings from comparing hotspot overlap, where mm10 gains were least likely to significantly overlap other hotspot types (Fig. 4,S9).

To confirm our findings and examine the GO terms themselves, we calculated the proportion of significant terms that were descendants (child terms) of a high-order parent term. Child terms were identified as statistically significant at a FDR < 0.05 based on a Fisher test using the classic algorithm. Additionally, we extracted the 10 highest ranked terms discovered using the Fisher test combined with 3 other algorithms designed to reduce false positives generated by the inheritance problem (described in methods) (Table S3-S6) (Alexa et al. 2006; Grossmann et al. 2007). Statistically significant terms for hg19 gain and loss mostly belonged to cellular processes, metabolic processes, single organism processes and biological regulation (Fig. 7). For mm10, DNA loss hotspots were enriched for similar terms, including developmental processes, which were particularly enriched in the mm10 genomic background (Fig S11). However, mm10 gain in the hg19 background was only enriched for a single term and in the mm10 background mm10 gain was not enriched for any terms. The difference in these results is consistent with how DNA gain and loss events in human and mouse associate with regions of varying gene density and biological activity (Fig. 5).

Interestingly, while the genomic distributions of each hotspot type differed, their associated significant GO terms were highly similar. This may be caused by genes that contribute to similar biological processes being tightly clustered and located within regions that consist of overlapping hotspot types. To determine if this was the case we compared non-redundant statistically significant child terms and gene annotations across each hotspot type (Fig S12). We found that the vast majority of genes annotated with significant GO terms were unique to a particular hotspot type. In contrast to this, the GO terms were more likely to be shared across hotspot types. This suggests that DNA gain and loss tend to associate with different genes that contribute to the same biological processes. Together our results show that particular biological processes are either prone to DNA gain or loss or are instead highly robust and able to withstand high levels of genomic turnover.

Discussion

Genome-wide DNA gain and loss dynamics

Estimating the total amount of DNA turnover across two separate lineages over a time span of approximately 90 million years is a challenging task (Hedges et al. 2006). After this divergence period as little as 40% of the extant human genome shares ancestry with mouse, suggesting that at least 60% has been turned over in either lineage. In order to understand gain and loss dynamics we must be able to correctly assign this non-aligning portion of the human genome as either human gain or mouse loss. Chinwalla et al. (2002) and Hardison et al. (2003) used an approach similar to our recent transposon based method. They used a set of lineage-specific transposons in human and mouse to identify regions of DNA gain. From this, the remaining non-aligning portion of one genome was assumed to be lost from the other. To confirm this approach, Chinwalla et al. (2002) checked to see if their inferred genome-wide rates of DNA loss were consistent with local estimates. They used the following equation;

$$G_E = G_A + G_G - G_L, \tag{2}$$

where G_E is the size of the extant genome, G_A is the size of the ancestral genome, G_G is the amount of lineage-specific genome gain and G_L is the amount of lineage-specific genome loss. For human and mouse they solved the equation for G_L where they estimated ancestral genome size within a range similar to the extant human genome size. This was chosen because it was similar to the average genome size for mammalian outgroup species. Estimates showed that DNA loss in mouse was almost double that of human, and consistent with the difference in the number of non-aligning non-recent transposon bases in each genome. While these estimates were consistent with expectations based on the assumption that non-aligning non-recent transposon regions were ancestral, their ancestral state remained unverified. Conversely, our ancestral based approach aimed to directly verify the ancestry status of non-aligning regions between human and mouse. This was achieved by using a wide variety of outgroup species alignments not available to Chinwalla et al. (2002) and Hardison et al. (2003) at the time of their analysis. In human, our results revealed that indeed many of the non-aligning non-recent transposon bases overlapped ancestral elements. However, approximately 168 Mb remained ambiguous (Table 2) which was more than double

the 5.8% of the total non-aligning human genome, the fraction of known ancestral bases not supported by ancestral elements (Table 1). As stated in the results, this discrepancy was most likely caused by incorrect identification of DNA gain events or misidentification of ancestral elements. It is important to realise that the ancestral element based approach has its limits, as orthologous sequences between species have the potential to diverge beyond recognition. This was the most likely reason that ancestral element detection in mouse was so much lower than in human, as the genome-wide substitution rate in mouse is approximately twice that of human.

An alternative way to verify the recent transposon based method was to use our estimated DNA loss rates to solve for G_A and to compare this to other estimates of ancestral genome sizes. After the mouse genome was completed many other mammalian genome projects also reached completion, allowing for the development of ancestral genome reconstruction techniques. While ancestral genome reconstruction is based on alignment it is much less susceptible to errors than our detection of ancestral elements. Instead of performing alignments directly between human or mouse and each individual outgroup species, it uses alignments between groups of more closely related species to build a phylogeny of ancestral states (Blanchette et al. 2004; Ma et al. 2006). Recently, Kim et al. (2017) estimated an ancestral euarchontoglires genome of 2.67 Gb in an analysis involving 19 placental mammals. Using equation 2 and solving for G_A with extant genome sizes from Table 1 and gain and loss rates calculated by the recent transposon method (Table 2), we get estimated ancestral genome sizes of 2.64 Gb and 2.66 Gb for human and mouse respectively. Together our findings in the context of various other methods support the use of recent transposons to analyse DNA gain and loss dynamics.

While the recent transposon method provides an accurate estimate of DNA gain and loss dynamics it is important to realise these estimates are only a lower bound on the the total amount of DNA turnover since divergence. This is because both our analysis and previous analyses relied heavily on the assumption of parsimonious genome evolution, where lineage-specific gain and loss patterns are based on the fewest possible evolutionary changes. Unfortunately, in our case the assumption of parsimonious genome evolution is likely to cause various events to be hidden. For example, if a particular region underwent lineage-specific DNA gain that was subsequently lost, both the gain and loss events will not be detected. Additionally, DNA loss occurring in both lineages at the same loci would also go undetected.

Depending on the frequency and magnitude of the above events we have likely underestimated the total amount of DNA gain and loss. A possible way to overcome this problem is to adopt model based approaches similar to those used in phylogenetic analyses. These approaches use a substitution model along with maximum likelihoods or Bayesian inference to allow for varying rates of evolution across lineages and sites (Yang and Rannala 2012). However, given our current lack of understanding of the non-coding portion of the genome such an approach for estimating DNA turnover is likely to yield highly questionable results.

Evolutionary impact of large scale DNA gain and loss

During genome evolution the spectrum of possible mutations is extremely broad, ranging from single nucleotide substitutions all the way up to Mb-sized rearrangements and translocations. Importantly, the genomic distribution of events at each level of the mutation spectrum is non-random and highly context-dependent. Moreover, the regional susceptibility and tolerance to a particular mutation type is a mixture of various genomic and epigenomic features and selective pressures (Makova and Hardison 2015). To understand the evolutionary impacts and trajectories of DNA gain and loss dynamics we analysed their genomic distributions in the context of various genomic features and biological processes.

In mammals synteny is highly conserved due to the frequent reuse of chromosome rearrangement breakpoints throughout their evolution (Murphy et al. 2005). Since chromosome rearrangement breakpoints were located outside of nets, many DNA gain and loss events went undetected (S1-S2). Instead, we most likely identified regions where gain and loss dynamics impacted on local architecture, such as the genomic distances between neighbouring genes or intron size. However, due to the difficulty in mapping DNA gain and loss events across large evolutionary time scales, the impact of DNA gain and loss at this scale remains largely unknown. Our strategy has therefore allowed us for the first time to measure regional variation in DNA gain and loss across genome structures that have been resistant to large structural rearrangements. Our results revealed that DNA gains and losses in human and mouse were associated with the same kinds of features; DNA gains were most associated with L1 accumulation in gene poor regions with low biological activity while DNA losses occurred mostly in highly active gene-rich regions. Previous analyses have shown that genome organisation between human and mouse is largely conserved, where lineage-specific L1s and SINEs

tend to accumulate in similar regions in different species (Buckley et al. 2017). Our results suggest that rather than certain types of events driving genome divergence, it is instead the rate at which each particular event type occurs that drives divergence. For example, mouse has a much higher deletion rate than human and a larger number of active L1s. This would suggest that particular regions in the mouse are growing or shrinking much more than in the human genome while their sequence composition remains similar. Alternatively, DNA gain rates were especially enriched on the X chromosome in both species with some degree of regional overlap (Fig. 4,S9). This is consistent with the high concentration of L1s that play a role in X inactivation (Chow et al. 2010).

Despite the amount of structural divergence between human and mouse, it is difficult to identify how much impact this might have on evolution at the level of phenotype. Interestingly, Human DNA gains and losses and mouse DNA losses all occurred near genes involved in fundamental cellular/metabolic processes. Because cellular/metabolic process genes likely evolved earlier in animals and probably have house keeping functions, their regulation is also likely highly conserved (Lowe et al. 2011). This suggests that for the most part the accumulation of DNA gains and losses have had little impact on phenotypic change. However, for some mouse DNA losses the case may be different, as in the mm10 genomic background they mostly occurred near genes involved in developmental processes. Developmental processes may be linked to traits that could have potentially undergone divergence, such as mouse-specific morphological characteristics. While this is an attractive idea, an analysis of regulatory element evolution shows that lineage-specific regulatory innovation for development occurred prior to human and mouse divergence (Lowe et al. 2011). Therefore, throughout mammalian evolution regulatory elements for development and cellular processes have likely remained intact while nearby DNA has been frequently turned over. Ultimately, given that we are able to detect little phenotypic impact where there are vast amounts of DNA turnover, our findings raise questions regarding the proportion of the human genome that is under selection and indeed 'functional'.

Topological associated domains (TADs) are a particular aspect of genome-organisation that may be affected by our detected DNA gains and losses. TADs are Mb-sized units of genome organisation that consist of highly self-interacting DNA. For example, two distant loci within a single TAD are much more likely to interact with each other than two loci that are near each other but happen to be located within different TADs (Dixon et al.

2012). Because TAD boundaries associate with other domain boundaries linked to gene regulation, such as LADs, they are often considered as distinct autonomously regulated regions (Sexton and Cavalli 2015). Since TADs are organised along a linear stretch of DNA, it is possible that their organisation is somewhat dependent on genomic distances between co-regulated features. This suggests that increased lineage-specific DNA gain and loss may cause TAD structures to diverge. One way this could happen is by removing TAD boundaries through deletion, which would subsequently cause adjacent TADs to merge (Hnisz et al. 2016). Alternatively, increases in the genomic distance between the edges of a single TAD could potentially promote the formation of a new boundary. These scenarios are more likely to have occurred in mouse rather than human, where DNA gain and loss in mouse is much more regionally clustered, ultimately causing larger deviations from regional gain and loss equilibrium. In vertebrates, *Hox* clusters are located between two adjacent TADs that most likely diverged from a single TAD leading to the evolution of the vertebrate *Hox* bipartite regulatory system (Acemel et al. 2016). This new TAD structure has made it possible for *Hox* genes to receive new inputs from distal enhancers contributing to the evolution of limb development and anteroposterior axis pattering (Lonfat and Duboule 2015). So while regulatory innovation at the level of individual elements may have slowed prior to human and mouse divergence, changes in TAD structure may cause ancestral enhancer elements to be co-opted in developmental processes driving lineage-specific phenotypic evolution.

Conclusion

There are four key points from our results. First, hot spots for DNA gains and losses occur in different compartments; loss hot spots in open chromatin/regulatory regions and gain hot spots in heterochromatin. Because DNA loss is caused by repair of DNA Double Stranded Breaks (DSB) (Gasior et al. 2006), this means that L1 ORF2p activity can both cause DNA gains and losses as a cause of DSB. However, this does not mean that gains and losses do not occur in the same regions. Second, mouse SINEs are strongly associated with DNA loss, indicating that losses in regulatory regions are accompanied by SINE insertions suggesting that there is extensive "churning" or turnover of sequences in these regions. The observed differences in associations between lineage-specific SINEs and gain and loss in mouse and human are likely due to differential expansion of LINEs vs SINEs in the two

lineages. Thus, regional/species specific variation in DNA gain and loss are primarily driven by clade specific/recent transposons interacting with open chromatin either in the male germ line, female germ line or early embryo. Third, the X chromosome is largely devoid of loss hot spots, but has many gain hot spots, consistent with a continuing selection for insertion of L1 elements required for X inactivation. Fourth, the observed autosomal divergence of gain and loss hot spot patterns in proximity to genes supports a model in which selection of altered developmental/regulatory mechanisms (based on GO term results) occurs as a result of transposon driven DNA gain and loss. This has implications for our views regarding the "functional" proportion of the genome that is under selection and contributing to phenotypic divergence.

Additional Files

Additional file 1 — Supplementary information

Competing interests

The authors declare that they have no competing interests.

Author's contributions

R.M.B., R.D.K., and D.L.A. designed research; R.M.B. performed research; and R.M.B., R.D.K., and D.L.A. wrote the paper.

Acknowledgements

We would like to thank Steve Pederson, Rick Tearle, Jonathan Henry Jacobsen, Lu Zeng and Zhipeng Qu for their helpful discussion throughout the research process and Catisha Coburn for help with editing the manuscript.

Availability of data and materials

References

Acemel, R. D., Tena, J. J., Irastorza-Azcarate, I., Marlétaz, F., Gómez-Marín, C., de la Calle-Mustienes, E., Bertrand, S., Diaz, S. G., Aldea, D., Aury, J.-M., et al. (2016).

A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate hox bimodal regulation. *Nature genetics*, 48(3):336–341.

Alexa, A. and Rahnenfuhrer, J. (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.26.0.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Berg, I. L., Neumann, R., Lam, K.-W. G., Sarbajna, S., Odenthal-Hesse, L., May, C. A., and Jeffreys, A. J. (2010). Prdm9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature genetics*, 42(10):859–863.

Berthelot, C., Muffato, M., Abecassis, J., and Crollius, H. R. (2015). The 3d organization of chromatin explains evolutionary fragile genomic regions. *Cell reports*, 10(11):1913–1924.

Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis*, 45(2):150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):1–36.

Blanchette, M., Green, E. D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research*, 14(12):2412–2423.

Brunschwig, H., Levi, L., Ben-David, E., Williams, R. W., Yakir, B., and Shifman, S. (2012). Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, 191(3):757–764.

Buckley, R. M., Kortschak, R. D., Raison, J. M., and Adelson, D. L. (2017). Similar evolutionary trajectories for retrotransposon accumulation in mammals. *bioRxiv*, page 091652.

Carlson, M. (2015). *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. R package version 3.2.2.

Carlson, M. (2016). *TxDb.Mmusculus.UCSC.mm10.knownGene: Annotation package for TxDb object(s)*. R package version 3.4.0.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Chow, J. C., Ciaudo, C., Fazzari, M. J., Mise, N., Servant, N., Glass, J. L., Attreed, M., Avner, P., Wutz, A., Barillot, E., et al. (2010). Line-1 activity in facultative heterochromatin formation during x chromosome inactivation. *Cell*, 141(6):956–969.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376.

Elliott, T. A. and Gregory, T. R. (2015). What's in a genome? the c-value enigma and the evolution of eukaryotic genome content. *Phil. Trans. R. Soc. B*, 370(1678):20140331.

ENCODE Project Consortium et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.

Fan, Y., Huang, Z.-Y., Cao, C.-C., Chen, C.-S., Chen, Y.-X., Fan, D.-D., He, J., Hou, H.-L., Hu, L., Hu, X.-T., et al. (2013). Genome of the chinese tree shrew. *Nature communications*, 4:1426.

Gasior, S. L., Preston, G., Hedges, D. J., Gilbert, N., Moran, J. V., and Deininger, P. L. (2007). Characterization of pre-insertion loci of de novo l1 insertions. *Gene*, 390(1):190–198.

Gasior, S. L., Wakeman, T. P., Xu, B., and Deininger, P. L. (2006). The human line-1 retrotransposon creates dna double-strand breaks. *Journal of molecular biology*, 357(5):1383–1393.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

Getis, A. and Ord, J. K. (1996). Local spatial statistics: an overview. *Spatial analysis: modelling in a GIS environment*, 374:261–277.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., and Warburton, P. E. (2007). Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS computational biology*, 3(7):e137.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? dna content, cell size, and the c-value enigma. *Biological reviews*, 76(1):65–101.

Gregory, T. R. (2005). The c-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Annals of botany*, 95(1):133–146.

Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2007). Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. *Bioinformatics*, 23(22):3024–3031.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948.

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome research*, 13(1):13–26.

Hedges, D. and Deininger, P. (2007). Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 616(1):46–59.

Hedges, S. B., Dudley, J., and Kumar, S. (2006). Timetree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., et al. (2006). The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, page aad9024.

International HapMap Consortium et al. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851.

Kamal, M., Xie, X., and Lander, E. S. (2006). A large family of ancient repeat elements in the human genome is under strong selection. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2740–2745.

Kapitonov, V. and Jurkal, J. (1996). The age of alu subfamilies. *Journal of molecular evolution*, 42(1):59–65.

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8):E1460–E1469.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, 100(20):11484–11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, 12(6):996–1006.

Kim, J., Farré, M., Auvil, L., Capitanu, B., Larkin, D. M., Ma, J., and Lewin, H. A. (2017). Reconstruction and evolutionary history of eutherian chromosomes. *Proceedings of the National Academy of Sciences*, 114(27):E5379–E5388.

Kirby, A., Kang, H. M., Wade, C. M., Cotsapas, C., Kostem, E., Han, B., Furlotte, N., Kang, E. Y., Rivas, M., Bogue, M. A., et al. (2010). Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics*, 185(3):1081–1095.

Kvikstad, E. M., Chiaromonte, F., and Makova, K. D. (2009). Ride the wavelet: a multiscale analysis of genomic contexts flanking small insertions and deletions. *Genome research*, 19(7):1153–1164.

Kvikstad, E. M., Tyekucheva, S., Chiaromonte, F., and Makova, K. D. (2007). A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS computational biology*, 3(9):e176.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome.

Laurie, S., Toll-Riera, M., Radó-Trilla, N., and Albà, M. M. (2012). Sequence shortening in the rodent ancestor. *Genome research*, 22(3):478–485.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9.

Lee, J., Hong, W.-y., Cho, M., Sim, M., Lee, D., Ko, Y., and Kim, J. (2016). Synteny portal: a web-based application portal for synteny block analysis. *Nucleic acids research*, 44(W1):W35–W40.

Lonfat, N. and Duboule, D. (2015). Structure, function and evolution of topologically associating domains (tads) athox loci. *FEBS letters*, 589(20PartA):2869–2876.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. *science*, 333(6045):1019–1024.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *science*, 302(5649):1401–1404.

Lynch, M. and Walsh, B. (2007). *The origins of genome architecture*, volume 98. Sinauer Associates Sunderland (MA).

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome research*, 16(12):1557–1565.

Makova, K. D. and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213–223.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584.

Murphy, W. J., Larkin, D. M., Everts-Van Der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617.

Nam, K. and Ellegren, H. (2012). Recombination drives vertebrate genome contraction. *PLoS genetics*, 8(5):e1002680.

Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2016). *RMySQL: Database Interface and 'MySQL' Driver for R.* R package version 0.10.8.

Pages, H. (2017). *BSgenome: Infrastructure for Biostrings-based genome data packages.* R package version 1.34.1.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R. M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4):603–613.

Petrov, D. A., Aminetzach, Y. T., Davis, J. C., Bensasson, D., and Hirsh, A. E. (2003). Size matters: non-ltr retrotransposable elements and ectopic recombination in drosophila. *Molecular biology and evolution*, 20(6):880–892.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Sexton, T. and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, 160(6):1049–1059.

Smit, A. F., Tóth, G., Riggs, A. D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of line-1 repetitive sequences. *Journal of molecular biology*, 246(3):401–417.

Smit, A. F. A., Hubley, R., and Green, P. (2013-2015). *RepeatMasker Open-4.0.* http://www.repeatmasker.org.

Team TBD (2014a). *BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19).* R package version 1.4.0.

Team TBD (2014b). *BSgenome.Mmusculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version mm10)*. R package version 1.4.0.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75.

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., et al. (2016). The ucsc genome browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634.

Vinogradov, A. E. and Anatskaya, O. V. (2006). Genome size and metabolic intensity in tetrapods: a tale of two lines. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1582):27–32.

Whitney, K. D. and Garland Jr, T. (2010). Did genetic drift drive increases in genome complexity? *PLoS genetics*, 6(8):e1001080.

Wickham, H. and Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. A., Gabriel, S. B., Reich, D., Donnelly, P., et al. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308(5718):107–111.

Wright, N. A., Gregory, T. R., and Witt, C. C. (2014). Metabolic 'engines' of flight drive genome size reduction in birds. In *Proc. R. Soc. B*, volume 281, page 20132780. The Royal Society.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H.-T., Nachman, M. W., Pialek, J., et al. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–655.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature reviews. Genetics*, 13(5):303.

Tables

Genomic regions	hg19	mm10
Sequenced genome	2897.0	2653.0
Gaps outside of nets	111.1	174.0
Non-RBH chains	306.1	293
Ancestral elements	1726.0	1021.0
Remaining chain-blocks	1014.3	994.4
Remaining chain-blocks ∩ ancestral elements (%)	94.2	85.2
Remaining chain-gaps	1465.8	1191.5

Table 1. Processing of net files. Sizes of genomic regions are measured in Mb unless otherwise specified.

hg19 chain-gaps				
		Ancestral element		
		hg19 gain	mm10 loss	**Total**
Recent transposon	hg19 gain	685.0	37.8	722.8
	mm10 loss	168.0	575.0	743.0
	Total	853.0	612.8	1465.8
mm10 chain-gaps				
		Ancestral element		
		mm10 gain	hg19 loss	**Total**
Recent transposon	mm10 gain	720.6	11.5	732.1
	hg19 loss	356.1	103.4	459.5
	Total	1076.7	114.9	1191.6

Table 2. hg19 and mm10 gap annotation. Chain-gaps were annotated using both the ancestral element and recent transposon method. Each number represents gap annotations in Mb.

Figures

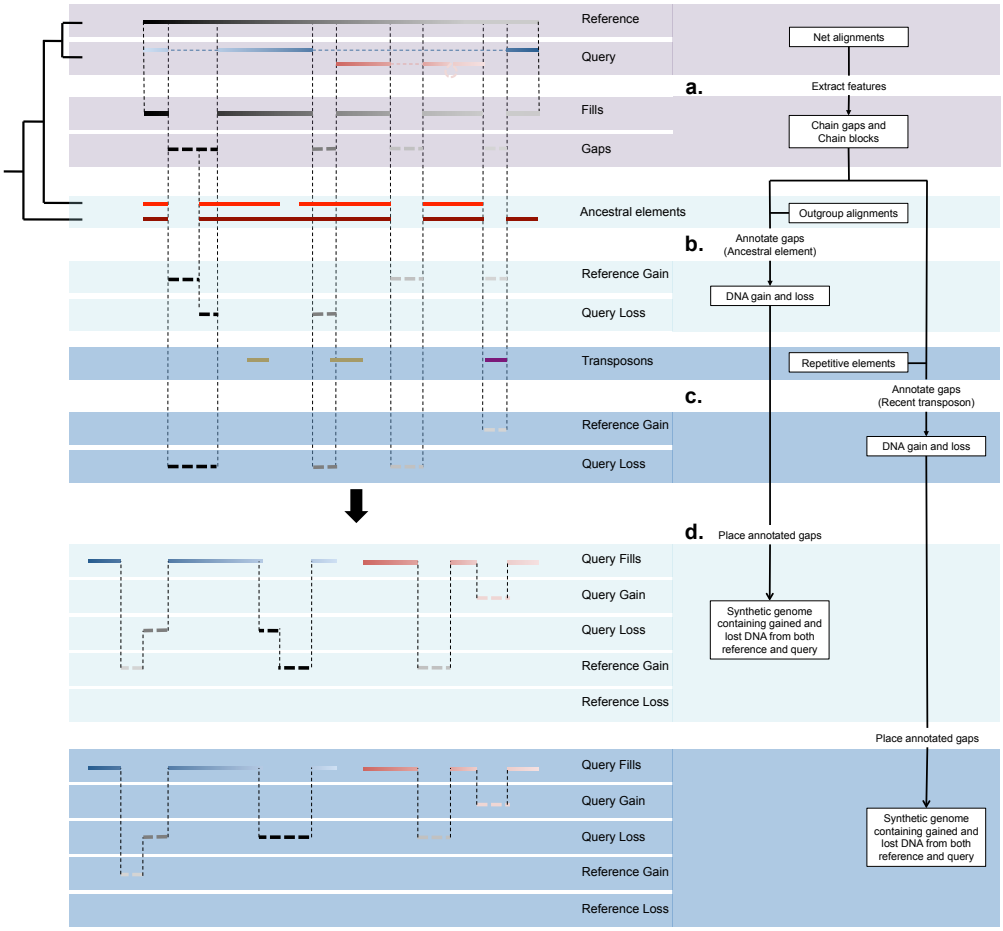


Figure 1. Detecting DNA gain and loss events between two species. Chain-gaps and chain-blocks are extracted from nets between reference and query (a). The resulting chain-gaps are essentially sequences from the reference genome that do not align to anything in the query genome. Chain-blocks are extracted from nets between reference and outgroup species as ancestral elements. Ancestral elements are then used to annotate chain-gaps as either gain or loss (b). Chain-gaps are annotated as query loss if they overlap ancestral elements or as reference gain if they do not. This is the ancestral element method for annotating gaps. The recent transposon method instead uses transposons classified as recent or ancestral to annotate gaps (c). Transposons are extracted from Repeat Masker files containing various classes of repetitive elements. Chain-gaps are annotated as reference gain if they overlap recent transposons or as query loss if they do not. After gaps are annotated they are placed within each genomic background creating a synthetic genome (d). Annotated chain-gaps are placed according to the edge coordinates of their adjacent chain-blocks within the same chain. Shown in the final two panels are chain-gaps extracted from the reference placed within the query genome. The different colours of the query chain-blocks show that gap annotations in the reference are placed on different chromosomes in the query. Differences in annotations are the results of conflicting information either resulting from incorrect identification of ancestral elements or recent transposons.

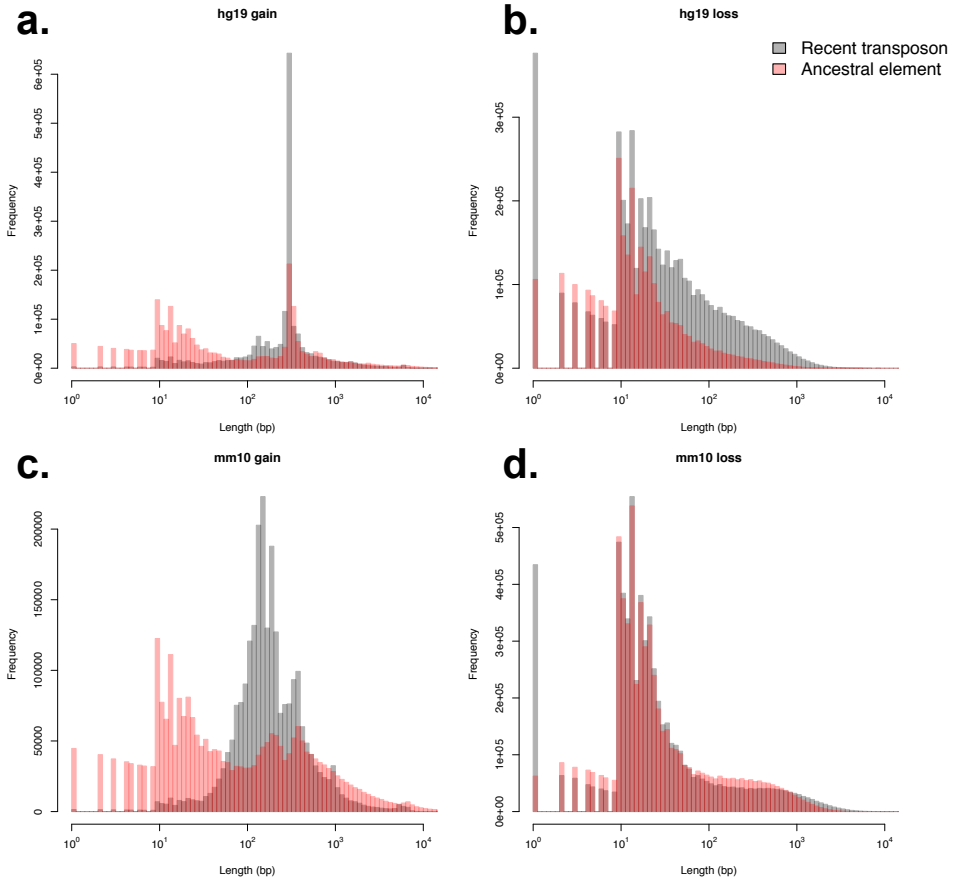


Figure 2. Length distributions of identified DNA gain and loss events. hg19 gain (a), mm10 gain (b), hg19 loss (c) and mm10 loss (d) events were identified using both the recent transposon and ancestral element method. Peaks for hg19 and mm10 gain, especially those detected by the recent transposon method, correspond to know lengths of transposon families.

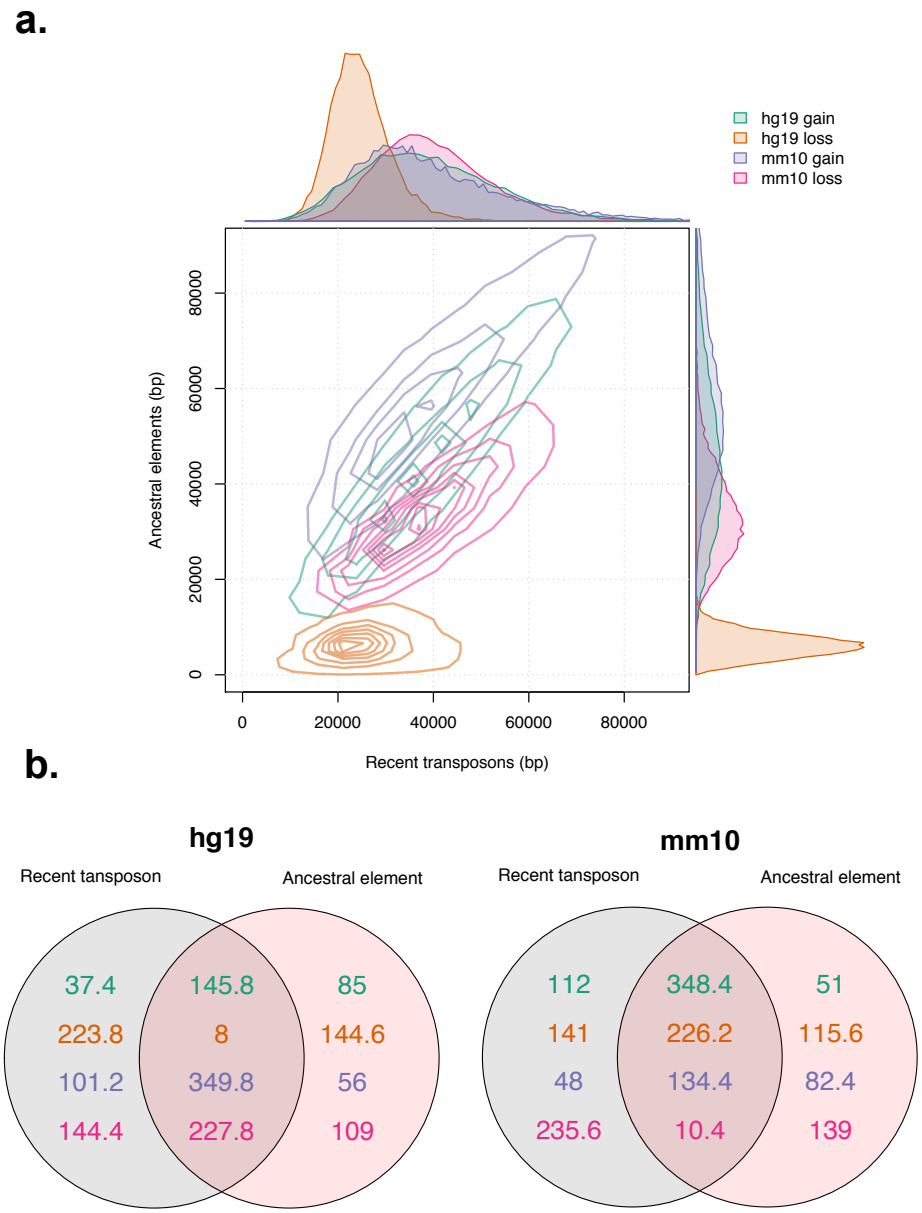
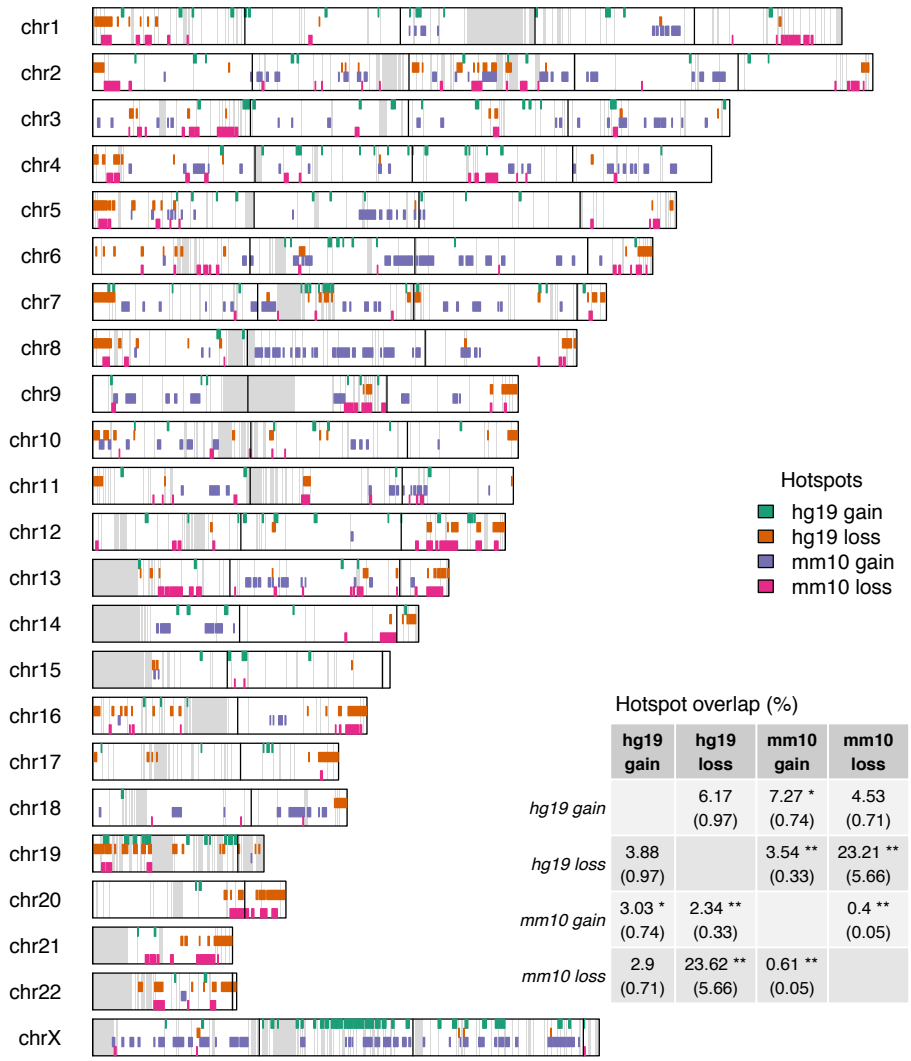


Figure 3. Comparison of gap annotation methods in binned synthetic genomes. Amount of DNA gain and loss per 200 kb in each bin for both hg19 and mm10 (**a**). For each gap annotation, contour lines begin at a 2D kernel density estimate of 2^{-10} and increase at regular intervals of 4^{-10}, except for hg19 which increase at regular intervals of 1.6^{-9}. Sizes of regions in Mb identified as hotspots for DNA gain or loss using the G_i^* statistic in each genome (**b**).



Hotspot overlap (%)

	hg19 gain	hg19 loss	mm10 gain	mm10 loss
hg19 gain		6.17 (0.97)	7.27 * (0.74)	4.53 (0.71)
hg19 loss	3.88 (0.97)		3.54 ** (0.33)	23.21 ** (5.66)
mm10 gain	3.03 * (0.74)	2.34 ** (0.33)		0.4 ** (0.05)
mm10 loss	2.9 (0.71)	23.62 ** (5.66)	0.61 ** (0.05)	

Figure 4. Genomic distribution of gain and loss hotspots for hg19 and mm10 plotted against hg19 synthetic genome. Grey regions indicate bins with <150 kb of RBH nets and black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss hotspots. The percentages were calculated using the hotspots labelled in each row as the denominator. '*' and '**' represent p-values below 0.05 and 0.01 respectively based on the Fisher statistic. The odds ratio for each fisher test is reported within the brackets. An odds ratio above 1 represents a positive association and an odds ratio below one represents a negative association.

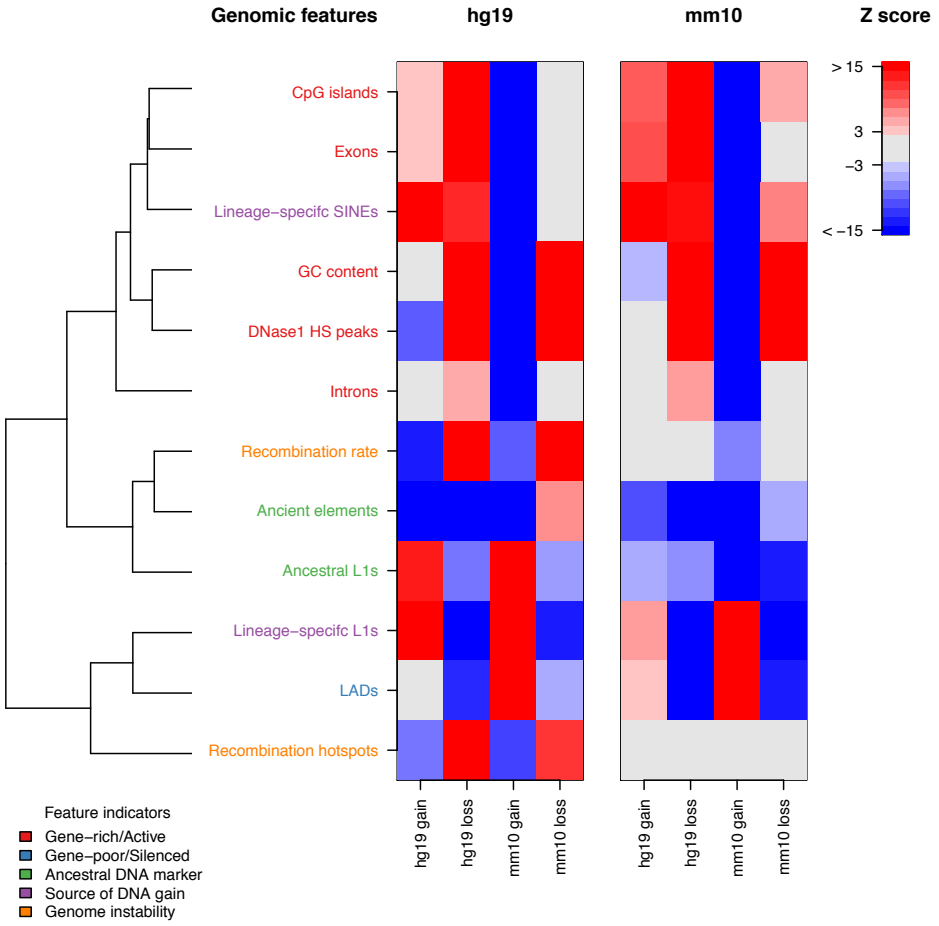


Figure 5. Association between genomic features and DNA gain or loss. Z scores are calculated using background distribution generated from 10000 permutations (methods). A positive association indicates that a particular gap annotation and genomic feature co-locate. Alternatively, a negative association indicates that the gap annotation and genomic feature occupy distinct genomic regions. DNaseI HS peaks (ENCODE Project Consortium et al. 2012), recombination hotspots (International HapMap Consortium et al. 2007; Brunschwig et al. 2012), LADs (Guelen et al. 2008; Peric-Hupkes et al. 2010), CpG islands (Tyner et al. 2016), gene annotations (Carlson 2015, 2016) and Retrotransposons (Smit et al. 2015) were measured in each as coverage per 200 kb. Recombination rates were measured as the mean bin-wise recombination rate (International HapMap Consortium et al. 2007; Brunschwig et al. 2012). GC content was measured as the proportion of G or C nucleotide residues in chain-blocks per bin (Team TBD 2014a,b). Genomic features are classified intro groups of feature indicators based on distinct aspects of genome biology they are known to associate with. The dendrogram represents spatial clustering of genomic features across both genomes,where two tightly clustered genomic features in the dendrogram are genomic features that tend to be co-located. The dendrogram was generated from a correlation matrix that consisted of pair-wise correlations between each feature across both binned genomes.

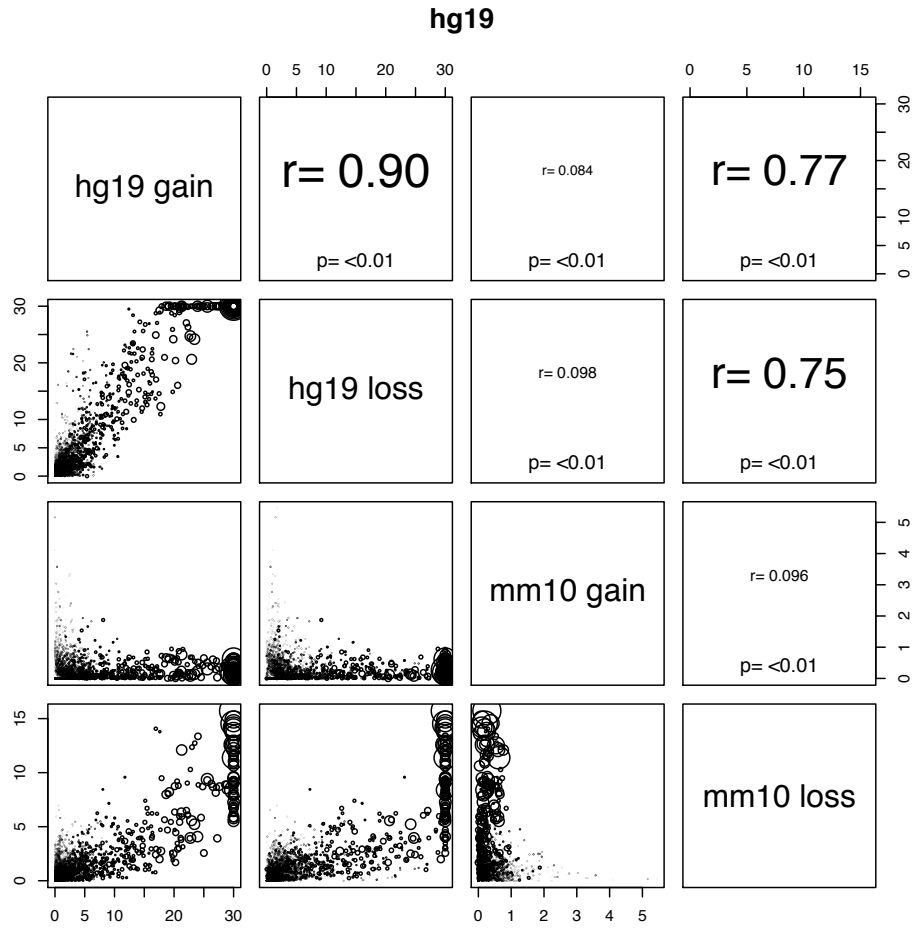


Figure 6. Over representation of biological process GO terms in gain and loss hotspots in hg19. The axes are marked according to -log10 P-values. The size of points represents the total number of annotations for each GO term.

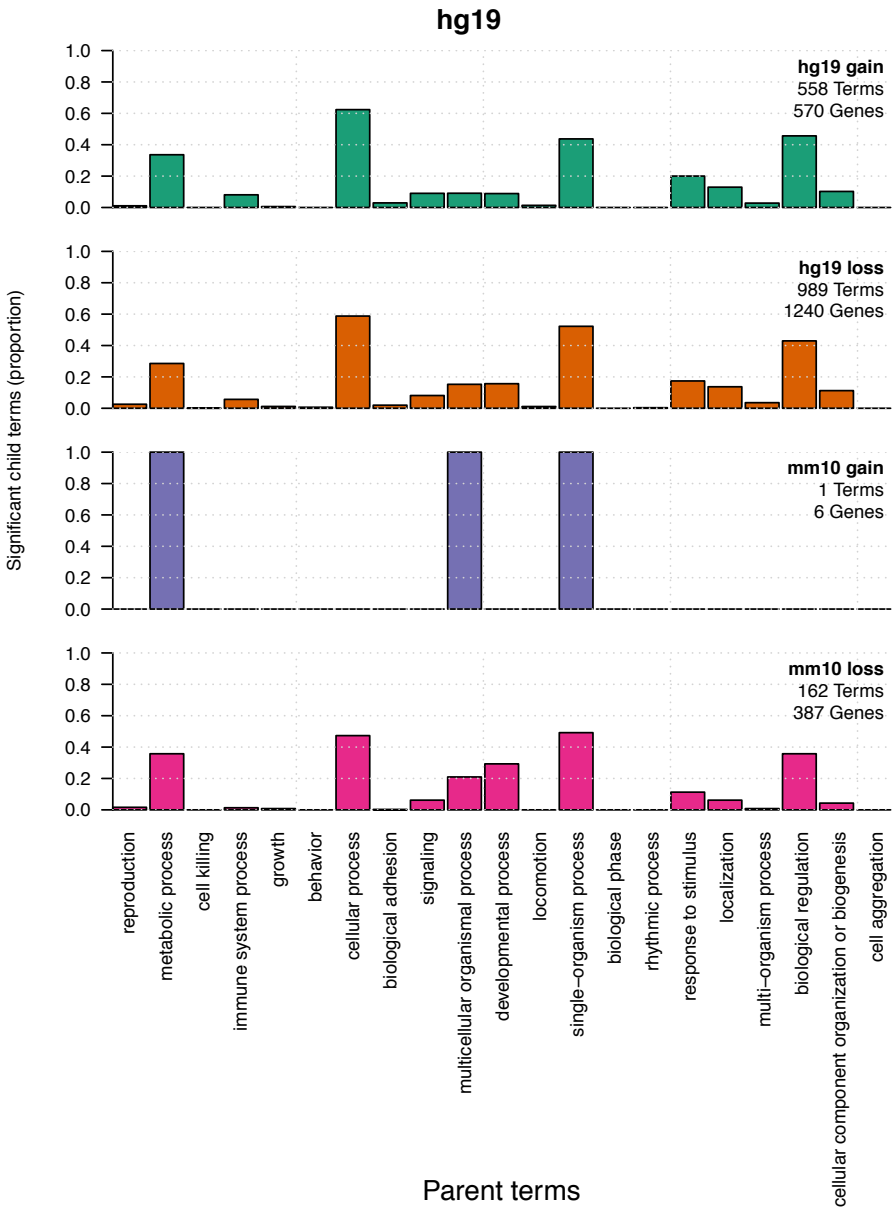


Figure 7. Significant biological process GO terms in hg19 background. Parent terms were the top level biological process GO terms while child terms were those beneath each parent term. Child terms were identified as significant at a FDR < 0.05 based on a Fisher test using the 'classic' algorithm. The Y axis represents the proportion of child GO terms that belong to each parant GO term. Proportions don't add up to 1 because some child GO terms are shared between parent GO terms. We have also shown the number of non-redundant GO terms and genes annotated with significant GO terms for each gap annotation.

Chapter 4

Bovine-specific transposable elements are associated with gene co-expression networks

Throughout this thesis I have focused specifically on species whose dominant retrotransposon is the L1. By comparing how similar element types accumulate in distinct evolutionary paths I was able to untangle the complex evolutionary relationships between LINE/SINE pairs and the genomes where they reside. However, due to various architectural similarities across the genomes I have studied, my findings provide a limited perspective. In this case the bovine genome is of central importance, its retrotransposon landscape is particularly distinct from many other placentals. This is because the dominant LINE in ruminants is the BovB element, a retrotransposon whose evolutionary origin in mammals is distinct from L1s. Throughout this chapter, I analyse the accumulation dynamics of retrotransposons in the bovine genome and how they associate with gene expression. I found that the divergent retrotransposon landscape of the bovine genome is strongly associated with several co-expression networks, revealing a link between genome organisation and gene expression. Finally, this chapter raises important questions regarding the impact of bovine-specific retrotransposons on genome evolution in ruminants.

Statement of Authorship

Title of Paper	Bovine-specific transposable elements are associated with gene co-expression networks
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Reuben M. Buckley, R. Daniel Kortschak and David L. Adelson. Bovine-specific transposable elements are associated with gene co-expression networks. 2017

Principal Author

Name of Principal Author (Candidate)	Reuben Buckley		
Contribution to the Paper	Processed data, performed analysis, prepared figures and wrote manuscript.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/08/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.		
Signature		Date	25/8/2017

Name of Co-Author	R. Daniel Kortschak		
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.		
Signature		Date	24/8/17

Bovine-specific transposable elements are associated with gene co-expression networks

Reuben M Buckley[1], R Daniel Kortschak[1], David L Adelson[1,*]

1 Department of Genetics and Evolution, The University of Adelaide, North Tce, 5005, Adelaide, Australia

*** david.adelson@adelaide.edu.au**

Keywords: Transposable elements, Gene expression networks, Bovine, Retrotransposon

Abstract

Transposable elements (TEs) are a major component of mammalian genome architecture and have the ability to alter and rewire gene regulatory networks. However, in most mammalian clades, TE composition is relatively similar and largely driven by the activity of long interspersed L1 (LINE_L1) elements. Due to this level of architectural conservation, it is difficult to use comparative genomics to calculate the true evolutionary impact caused by TE accumulation. To provide a broader perspective to the field of TE mediated mammalian genome evolution, we analysed the distribution of TEs in the bovine genome and their association with gene expression. Importantly, TE evolution in the bovine genome is distinct from other mammals as it is largely driven by the LINE_BovB element, a TE class only found in a few mammalian clades outside of ruminants. We characterised co-expression modules across 16 tissue samples using weighted gene co-expression network analysis. Next, we analysed their TE associations by measuring the density of TEs surrounding their member genes. Across all TEs, we found that gene co-expression modules were only associated with bovine-specific groups. We also found that based on module expression activity, bovine-specific TE associated modules clustered separately from gene-rich/TE-poor modules. This finding echoed the genomic distribution of individual TE families, where bovine-specifc TEs accumulated in gene-poor regions. Together our results show that bovine gene expression is strongly associated with genome structure. Although, large-scale TE accumulation in human is not usually associated with gene expression evolution, there are key differences in bovine-specific TE accumulation dynamics that provide new scope for further analysis. The fact that these elements accumulate in regions that are usually associated with relaxed selection for gene regulation means that bovine-specific TEs are more primed for regulatory innovation than TEs in other mammalian species.

Introduction

The surrounding genome architecture of a gene is strongly linked to its expression, regulation and function, yet many studies analyse these factors independently. One major component of genome organisation linked to gene expression is transposable elements (TEs), interspersed self-replicating mobile DNA elements. Their insertion has the potential to alter nearby expression and their accumulation patterns are known to vary according to the family they belong to (Chuong et al. 2017). Across placentals, TEs occupy almost 50% of their genome sequence with most of their evolution occurring after they diverged from a common ancestor (Lander et al. 2001; Chinwalla et al. 2002; Adelson et al. 2009; Kapusta et al. 2017). This means that many sequences repeated throughout primates are absent from ruminants and vice versa, leading to massively divergent genome composition between these taxa. While there are many instances where a particular TE has been identified as a gene regulatory element, an understanding of how their broader accumulation patterns associate with gene expression is lacking.

One of the main techniques for studying TE and gene expression evolution is comparative genomics, which uses observational data to gain insight into the dynamic processes that lead to divergent forms. One of the limitations of this approach is the availability of high quality sequenced genomes. This is why most comparative genome-wide analyses that focus on gene expression and regulation are performed between human and mouse (Yue et al. 2014). While this work has yielded large insights into mammalian genome evolution, certain mouse-specific genomic factors limit further discovery of genome evolutionary dynamics. For example, the genomes in the rodent lineage are fast evolving and frequently undergo genomic rearrangements (Chinwalla et al. 2002). This makes it difficult to perform basic alignments between non-coding regions and identify regions that may have shared ancestry. Additionally, different sources of structural evolution such as TEs, deletions and rearrangements are likely confounded, making it difficult to measure their individual evolutionary impact on gene expression.

An alternative and often overlooked clade that could be helpful for understanding the evolutionary impact of TEs is the ruminants. Ruminant chromosomes have remained largely intact since divergence from human and have also been subjected to a much lower substitution rate than mice (The Bovine Genome Sequencing and Analysis Consortium et al.

2009). However, more importantly, ruminants have a fundamentally divergent TE landscape from other mammalian lineages at similar phylogenetic distances (The Bovine Genome Sequencing and Analysis Consortium et al. 2009; Adelson et al. 2009). The major TE classes in primates, rodents, carnivores, perissodactyls and non-ruminant cetartiodactyls are long interspersed L1 elements (LINE_L1s) and their associated clade-specific short interspersed elements (SINEs) (Ivancevic et al. 2016, 2017). In ruminants the major TE class is instead LINE_BovB which was introduce to the lineage through an ancient horizontal transfer event and now occupies over 10% of the bovine genome (Adelson et al. 2009; Walsh et al. 2013). This is important, as the accumulation patterns of LINE_L1s and their associated SINEs in different species follow similar evolutionary trajectories (Buckley et al. 2017). Since LINE_BovBs have an evolutionary origin and insertion mechanisms distinct from LINE_L1s, ruminants provide a unique window in which to identify general principals regarding the evolutionary impacts of TEs.

To begin to untangle the complex relationships between TEs and their role in mammalian genome evolution, we used a weighted gene co-expression network analysis (WGCNA) approach to explore the association between TE accumulation and gene expression. This approach was applied to the bovine genome using an RNA-seq dataset comprised of 16 different tissues sampled from L1 Dominette 01449, the individual from which the bovine reference sequence was obtained (The Bovine Genome Sequencing and Analysis Consortium et al. 2009). We found that TEs most associated with gene expression were usually bovine specific, indicating their accumulation may have had an evolutionary impact on the bovine transcriptome. Additionally, these bovine specific TE associated modules also clustered separately from co-expression modules enriched with genes. Based on the accumulation of bovine specific TEs in gene poor regions our results highlight the importance of genome architecture on the evolution of gene expression. Ultimately, our analysis and findings help to bridge the gap between a gene's activity and its broader genomic context.

Materials and methods

Obtaining RNA-seq data

Tissue samples were collected from L1 Dominette 01449 and mRNA Libraries were prepared using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego) (Taylor et al. 2016).

RNA was sequenced using paired-end 100 bp reads with an approximate 175bp insert at the Beijing Genomics Institute (`http://bgi-international.com`). Three technical replicates were performed for each individual sample and were distributed across three separate sequencing lanes. Sequence data was provided by Jeremy F. Taylor at the University of Missouri and can be found in the short read archive under the following accession: SRP063069. Further information regarding individual sequencing runs can be found in additional file 2.

Data processing and RNA-seq normalisation

We measured read quality using FastQC (Andrew 2010) and based on these results we trimmed reads using the FASTX-Toolkit (Hannon lab 2010). For each read we trimmed positions 1-15 and 95-100; any remaining read that was less than 60 bp in length was discarded. Because read-1 and read-2 of each read-pair was processed separately their ordering within our FASTQ files no longer reflected their pairing. To correct this we paired reads using their read identifiers and removed all reads whose mate-pair was discarded from either the read-1 or read-2 FASTQ file. Next, read-pairs were mapped to the UMD3.1 bovine reference assembly using the program Subread and only accepted uniquely mapped reads (Liao et al. 2013b). The reference assembly was obtained from the Ensembl database release 74 (Flicek et al. 2014) (`ftp://ftp.ensembl.org/pub/release-74/fasta/bos\_taurus/dna/Bos\_taurus.UMD3.1.74.dna.toplevel.fa.gz`). After read mapping we counted the number of read-pairs mapped to each gene annotation across all of our datasets using the tool featureCounts (Liao et al. 2013a). For this we used gene annotations also obtained from Ensembl database release 74 (`ftp://ftp.ensembl.org/pub/release-74/gtf/bos_taurus/Bos_taurus.UMD3.1.74.gtf.gz`). Next, we carried out FPKM normalisation on our read-pair counts and discarded genes that were not located on assembled chromosomes or had expression levels equal to zero in at least one dataset.

Identifying co-expression modules

We identified co-expression modules using the WGCNA package (Langfelder and Horvath 2008, 2012). For WGCNA, we stabilised gene expression variance by using $log_2(x+1)$ to transform our data, where x was our FPKM expression levels. Next, we built multiple co-expression networks based on gene expression correlation between gene pairs. Based on WGCNA guidelines we chose the signed network option and fitted our data to a scale-

free topology model. Our co-expression modules were detected using automatic network construction in a block-wise manner. The settings we used were a minimum module size of 30 genes and a dendrogram cut height for module merging of 0.25.

Identification and classification of TEs

TE coordinates were identified in the UMD3.1 bovine assembly using Censor combined with a repbase library that consisted of TEs found across mammals (Kohany et al. 2006). We created a series of TE groups defined on the basis of TE class and period of activity, which was either ancestral (found throughout mammals) or bovine-specific (found only in ruminants). Our ancestral TE groups were named ERV_anc, LINE_L1_anc, LINE_L2 and SINE_MIR. Our bovine-specific TE groups were named ERV_BT, LINE_BovB, LINE_L1_BT, SINE_BOVA2 and SINE_BOVTA. Individual TEs were placed into groups based on a series of regular expressions that matched identifiers for TE class and period of activity found in their family names. A table of the regular expressions that were used can be found in the supplementary information (Table S1).

Analysing the genomic distribution of TEs

To correctly analyse the genomic distribtuion of TEs we need to segment the bovine genome at an appropriate bin-size. To do this, the bovine genome was segmented at multiple bin-sizes and the bin-wise coverage level of each TE group was tallied and normalised by the number of sequenced bases in each bin. For each of these segmented genomes we calculated the spatial autocorrelation between genomic bins and their downstream neighbour. From this we selected a bin-size of 250 kb. Next, we analysed TE distributions by calculating Spearman's rank correlation for each pair of TE groups. Gene number was included in this analysis to help provide context and anchor our results to a well established feature of genome organisation across mammals.

Detecting TE associated co-expression modules

Each gene was assigned a TE score for each TE group. Scores were assigned based on the TE content of the genomic bin overlapping each gene's transcription start site. To make these scores comparable across TE groups we standardised them by calculating their Z scores. Using these scores we were able to detect statistically significant TE associated co-expression

modules using a permutation based approach. Gene module membership was shuffled across our genes and we calculated each module's resampled mean TE score. This process was repeated 10,000 times to generate a background distribution of resampled mean TE scores for each module. These resampled TE score distributions were then used to convert our observed mean module TE scores into Z scores and determine the association strength between each module's member genes and each TE group. Based on multiple testing of 420 modules across 10 TE/gene groups, only those module associations with a Z-score outside the range of -3.65 – 3.65 were considered statistically significant. This is because a Z score > 3.65 is equal to a false discovery rate (FDR) < 0.05

Gene Ontology term enrichment

For specific co-expression modules we calculated Gene ontology (GO) term enrichment for biological process terms (BP). We used topGO to calculate P-values for each term based on the Fisher test and defined statistical significance at a FDR < 0.05 (Alexa and Rahnenfuhrer 2016).

Software used for data analysis

To analyse our data we used the following packages within the R environment(R Core Team 2016): Genomic Ranges (Lawrence et al. 2013), dplyr (Wickham and Francois 2015), org.Bt.eg.db (Carlson 2016b), GO.db (Carlson 2016a) and Bioconductor (Huber et al. 2015).

Results

Co-expression module detection

After initial processing and normalisation of RNA-seq data (methods), our dataset included expression levels for 12280 Ensembl annotated genes across 48 sequencing runs from a total of 16 tissue samples. To determine if there was any sample/run-specific bias, we performed hierarchical clustering on our entire dataset of sequencing runs (Fig. 1). We compared this clustering pattern to run specific sequencing and mapping statistics generated by Subread (Liao et al. 2013b). Our results showed that individual sequencing runs clustered according to their corresponding tissue samples. Interestingly, the number of raw reads per run varied

according to sequencing lane and machine ID, however this variation was factored out by our normalisation procedure. Together our results indicate that the vast majority of variation across our dataset was biological rather than technical.

Next, we applied WGCNA to our log normalised gene expression dataset to detect co-expression modules (methods). We chose a soft threshold of 14 which gave low levels of connectivity and a scale-free topology model fit of almost 0.8 (Fig. S1). Since, our dataset was quite large we found it easier to detect modules across 3 separate blocks (Fig. S2). Using this approach we identified a total of 43 co-expression modules with gene membership ranging between 30 and 3000 genes (Fig. S3).

The bovine TE landscape

To characterise the bovine TE landscape and measure its association with gene expression, we placed TEs from various families into 9 distinct TE groups using their family identifiers (Table S1) (methods). Based on similarity to TE consensus sequences, our grouping strategy successfully identified both bovine-specific and ancestral TEs (Fig. 2a). The percent similarity interquartile range for bovine-specific TE groups was $> 80\%$, indicating these elements had only recently diverged from a common ancestor. In contrast, the percent similarity interquartile range for ancestral TE groups was $< 80\%$, indicating that these elements have been diverging for a much longer period of time than bovine-specific TEs.

To capture the bovine TE landscape, we segmented the bovine genome into equally sized genomic bins. It is worth noting that for many analyses on binned genomes, bin-size is an important factor. An example of this, is the genomic distribution of TEs and various other genomic features within the horse genome. Adelson et al. (2010) showed that changes in bin-size caused genomic spatial associations between some of these features to either strengthen, weaken or even change sign. This indicates that inappropriate bin-size choice may result in unpredictable and misleading outcomes. To ensure that we chose the optimal scale for our analysis, we explored how genomic spatial autocorrelation changed as a function of bin-size (methods). The reason we used spatial autocorrelation is that it reflects the degree to which a particular feature is locally clustered across an entire genome or landscape (Moran 1950). At a bin-size where there are high levels of genome-wide spatial clustering of a particular feature, shifts in that feature's local density between neighbouring bins is

gradual. This means that we are capturing the scale at which biologically relevant forms and patterns begin to emerge regarding that feature's genomic distribution. Alternatively, at a smaller bin-size where there are lower levels of spatial clustering, changes in feature density between neighbouring bins will appear more stochastic. Therefore, we chose a bin-size of 250 kb as it provided the best compromise between genomic resolution and spatial clustering across the majority of our TE groups (Fig. S4).

After selecting the appropriate bin-size for genome segmentation, we used correlation analysis to characterise the bovine TE landscape (methods). We found that bovine-specific TEs accumulated in gene-poor regions, where ancestral LINE_L2 and SINE_MIR TE groups accumulated more in gene-rich regions (Fig. 1b). LINE_L2s and SINE_MIRs are an inactive ancestral LINE/SINE pair, however their accumulation patterns have remained strongly conserved across distantly related species (Adelson et al. 2009, 2010). In contrast to this, bovine-specific LINE_L1s and LINE_BovBs are actively replicating TE groups that have both caused large levels of lineage-specific divergence. Since L1s are found throughout mammals and tend to accumulate in similar regions, they have been characterised as independent agents that drive lineage-specific mammalian evolution along similar trajectories (Buckley et al. 2017). Interestingly, while LINE_BovBs are compositionally distinct from LINE_L1s, they both share similar accumulation patterns (Fig 2b). As a result, the LINE content of the bovine genome appears to be positionally conserved. However, because LINE_BovBs are absent from the genomes of many other mammalian species, the bovine genome appears to be compositionally distinct.

Potentially one of the most outstanding features of the bovine TE landscape is the distribution of bovine-specific SINEs. Unlike human and mouse, whose lineage-specific SINEs accumulate in gene-rich regions (Lander et al. 2001; Chinwalla et al. 2002), bovine-specific SINEs accumulate in gene-poor regions (Fig. 2). This may be because bovine-specific SINEs are mobilised by LINE_BovB replication machinery, as opposed to the LINE_L1 replication machinery that mobilises human- and mouse-specific SINEs (Ohshima and Okada 2005). Moreover, the genomic distribution of the individual bovine-specific SINE groups, SINE_BOVA2 and SINE_BOVTA, is also quite quite complex. For example, SINE_BOVA2 and SINE_BOVTA TEs occasionally co-locate, however their overall spatial correlation patterns with other TE groups are quite distinct from each other. Since LINE_BovBs are the driving force behind bovine-specific SINE activity, they have not only contributed to the

compositional divergence of bovine TE genome architecture, but they are also major drivers of the positional divergence of bovine TE genome architecture. This divergence in genome structure may provide some of the raw material for bovine-specific gene regulatory and gene expression evolution.

Bovine-specific TEs are associated with co-expression modules

After characterising the bovine TE landscape, we analysed gene expression patterns based on their surrounding TE content. We did this by identifying co-expression modules whose member genes were located in regions enriched for TEs. This was done using a permutation approach based on 10,000 iterations (methods). Out of our 42 modules we identified a total of 10 that showed some kind of statistically significant TE association for at least 1 TE group (Fig 3a).

Our results showed that bovine-specific TEs were the only TE groups to be positively associated with any of our co-expression modules. The TE group that had the most associations with co-expression modules was SINE_BOVA2. Their frequent over representation in co-expression modules suggests that depending on their ability to alter gene expression, they may be responsible for a large amount of bovine-specific evolution. This kind of accumulation near genes in specific co-expression modules along with high sequence similarity, is consistent with SINE BOVA2s providing transcription factor binding sites (TFBSs). As TFBSs, BOVA2 elements would exist as fundamental components for the wiring of specific gene regulatory networks. Consistent with this idea, previous analyses have identified a SINE_BOVA2 upstream of the TP53 gene in many non-domesticated bovids. TFBS analysis of these elements revealed that they carry unique binding sites for transcription factors that likely regulate TP53 expression and mammary involution (Dekel et al. 2015). Another role for SINE_BOVA2s in gene regulation, is as targets for micro RNAs. A core motif in the SINE_BOVA2 element has been shown to act as a target site for several bovine-specific micro RNAs. In addition, SINE_BOVA2s have been found in the 3' untranslated regions of mRNAs involved in cell growth and differentiation during the immune response (Damiani et al. 2008).

Other bovine-specific TEs that were positively associated with co-expression modules included LINE_BovB and SINE_BOVTA. Potential molecular roles for these elements are

lacking in the literature, however based on discovery in other mammals they still may play an important role in gene expression (Chuong et al. 2017). Interestingly, LINE_BovB occasionally shares module association with SINE_BOVA2s, whereas SINE_BOVTAs are positively associated with co-expression modules in a more exclusive manner. This is likely linked to the underlying genomic distribution of these elements discussed above. Our results also showed that there were 2 modules that were positively associated with genes that clustered with a third module that was negatively associated with bovine-specific TEs. This suggests that gene rich areas of the genome contain genes with similar expression patterns and supports the idea that mammalian genomes are organised according to roles/regulation of particular genes.

Next, for each TE associated module we identified statistically enriched BP GO terms (methods). To gain an overview of the functional attributes for each module we tallied our statistically enriched terms according to their top level ancestor BP GO terms (Fig 3b). Across our dataset, GO terms were mostly associated with the gene-rich green module which contained a wide range of biological processes. Other modules that associated with GO terms were the turquoise, lightyellow and black modules, which were mostly associated with metabolic and cellular processes. It is worth noting that the R package 'org.Bt.eg.db' only contained approximately 3000 genes with annotated BP GO terms (Carlson 2016b). This means that out of our dataset of approximately 12000 genes, it is likely that many modules may not contain any genes with an annotated GO term. Therefore, absence of GO term association with a particular module does not mean that the roles of genes within that module do not correspond to a particular biological process.

Another way to analyse our TE associated modules is to measure their eigengene expression across our tissue samples (Fig 3c). Intuitively, eigengenes can be thought of as a single gene whose expression is representative of an entire co-expression module (Langfelder and Horvath 2007). From clustering our modules based on eigengene expression, we observed that module TE enrichment strongly associated with their expression patterns; gene-rich/TE-poor modules clustered separately from our bovine-specific TE enriched modules. Since genes and bovine-specific TEs accumulate in distinct genomic regions (Fig. 2b), our results are consistent with genome organisation where genes involved in similar processes are clustered in similar regions.

Discussion

Collectively, our results present a complex relationship between gene expression and genomic TE distributions. We identified specific groups of similarly expressed genes that tend to be located in regions with a high concentration of bovine-specific TEs. This observation can be explained by two separate hypotheses. The first is the TE regulatory hypothesis, where bovine-specific TEs provide some sort of regulatory signal required for coordinated gene expression of specific gene sets. This suggests that the introduction of LINE_BovBs into a ruminant ancestor is a major evolutionary event that may have driven the evolution and diversification of the ruminant lineage. The second is the genome organisation hypothesis, where genes are organised according to the processes they are involved in and the functions they perform. In this case TEs are enriched in co-expression modules because their member genes are located in regions where TEs tend to accumulate. While there are plenty of examples where TEs have been exapted into regulatory elements, analysis of their wider genomic distributions tends to support the genome organisation hypothesis. In hominids, expression divergence between human and chimpanzee ortholog gene pairs associates with recent ancestral TE insertions more than lineage-specific TE insertions. This suggests that TE accumulation occurs near genes with relaxed selection and has little impact on gene expression itself (Warnefors et al. 2010). Moreover, TE accumulation in hominids is most associated with germ-line expressed genes, indicating that TEs sometimes preferentially accumulate near genes with similar expression behaviour (Warnefors et al. 2010). These findings are further supported by relatively rare TE exaptation linked to pleiotropic constraints regarding regulatory element innovation (Nikolov and Tsiantis 2017). However, it is important to realise that these findings occurred across a very small groups of species with TE accumulation dynamics that are very different to the bovine. The majority of lineage-specific TE insertions in human are due to SINE_*Alu* elements and have occurred in gene-rich open chromatin regions. Since these regions tend to be occupied by house keeping genes that have highly conserved expression profiles, SINE_*Alu* insertions that impacted on gene expression would likely be selected against (Buckley et al. 2017). Conversely, bovine-specific SINEs instead accumulate in gene-poor regions, where genes tend to have more tissue-specific expression patterns. In these regions gene regulation is under relatively relaxed selection and more primed toward regulatory innovation, leading to the evolution of lineage-defining traits

(Nikolov and Tsiantis 2017).

In conclusion, we identified various co-expression modules that showed statistically significant positive associations with bovine-specific TE groups. We found that based on eigengene expression, these TE associated modules clustered separately from gene-rich/TE-poor modules. This was similar to the genome-wide distributions of each TE group, where bovine-specific TEs accumulated in gene-poor regions. While it was difficult to ascertain the actual evolutionary impact of bovine-specific TEs on gene expression, our results show that genome organisation is strongly associated with gene behaviour. Importantly, we identified key differences between the bovine TE landscape and the TE landscapes of most other mammals that have been previously analysed. These differences provide scope for further analysis into gene regulation from bovine-specific TEs and highlight the importance of using diverse species to study the evolution of genome architecture in mammals.

Additional Files

Additional file 1 — Supplementary information

Supplementary figures and tables

Additional file 2 — BovineRNASeqSupInfo.xlsx

Supplementary information for each individual sequencing run used in the analysis.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

R.M.B., R.D.K., and D.L.A. designed research; R.M.B. performed research; and R.M.B., R.D.K., and D.L.A. wrote the paper.

Acknowledgements

The authors would like to thank Zhipeng Qu for helpful discussion and advice, and Jeremy F. Taylor for providing the RNA-seq data.

Availability of data and materials 327

References

Adelson, D., Raison, J., Garber, M., and Edgar, R. (2010). Interspersed repeats in the horse (equus caballus); spatial correlations highlight conserved chromosomal domains. *Animal genetics*, 41(s2):91–99.

Adelson, D. L., Raison, J. M., and Edgar, R. C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences*, 106(31):12855–12860.

Alexa, A. and Rahnenfuhrer, J. (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.26.0.

Andrew, S. (2010). Fastqc: a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`.

Buckley, R. M., Kortschak, R. D., Raison, J. M., and Adelson, D. L. (2017). Similar evolutionary trajectories for retrotransposon accumulation in mammals. *bioRxiv*, page 091652.

Carlson, M. (2016a). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.4.0.

Carlson, M. (2016b). *org.Bt.eg.db: Genome wide annotation for Bovine*. R package version 3.4.0.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature reviews. Genetics*, 18(2):71.

Damiani, G., Florio, S., Panelli, S., Capelli, E., and Cuccia, M. (2008). The bov-a2 retroelement played a crucial role in the evolution of ruminants. *Rivista di biologia*, 101(3):375.

Dekel, Y., Machluf, Y., Ben-Dor, S., Yifa, O., Stoler, A., Ben-Shlomo, I., and Bercovich, D. (2015). Dispersal of an ancient retroposon in the tp53 promoter of bovidae: phylogeny, novel mechanisms, and potential implications for cow milk persistency. *BMC genomics*, 16(1):53.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic acids research*, 42(Database issue):D749–55.

Hannon lab (2010). Fastx-toolkit: Fastq/a short-reads pre-processing tools. `http://hannonlab.cshl.edu/fastx_toolkit/index.html`.

Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., Bravo, C., H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, D., K., Irizarry, A., R., Lawrence, M., Love, I., M., MacDonald, J., Obenchain, V., Ole's, K., A., Pag'es, H., Reyes, A., Shannon, P., Smyth, K., G., Tenenbaum, D., Waldron, L., Morgan, and M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Ivancevic, A., Kortschak, D., Bertozzi, T., and Adelson, D. (2017). Re-evaluating inheritance in genome evolution: widespread transfer of lines between species. *bioRxiv*, page 106914.

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., and Adelson, D. L. (2016). Lines between species: Evolutionary dynamics of line-1 retrotransposons across the eukaryotic tree of life. *Genome biology and evolution*, 8(11):3301–3322.

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8):E1460–E1469.

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. *BMC bioinformatics*, 7(1):474.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome.

Langfelder, P. and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1(1):54.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Langfelder, P. and Horvath, S. (2012). Fast r functions for robust correlations and hierarchical clustering. *Journal of statistical software*, 46(11).

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9.

Liao, Y., Smyth, G. K., and Shi, W. (2013a). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Liao, Y., Smyth, G. K., and Shi, W. (2013b). The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108–e108.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Nikolov, L. A. and Tsiantis, M. (2017). Using mustard genomes to explore the genetic basis of evolutionary change. *Current Opinion in Plant Biology*, 36:119–128.

Ohshima, K. and Okada, N. (2005). Sines and lines: symbionts of eukaryotic genomes with a common tail. *Cytogenetic and genome research*, 110(1-4):475–490.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Taylor, J. F., Whitacre, L. K., Hoff, J. L., Tizioto, P. C., Kim, J., Decker, J. E., and Schnabel, R. D. (2016). Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. *Genetics Selection Evolution*, 48(1):59.

The Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G., Tellam, R. L., Worley, K. C., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324(5926):522–528.

Walsh, A. M., Kortschak, R. D., Gardner, M. G., Bertozzi, T., and Adelson, D. L. (2013). Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences*, 110(3):1012–1016.

Warnefors, M., Pereira, V., and Eyre-Walker, A. (2010). Transposable elements: insertion pattern and impact on gene expression evolution in hominids. *Molecular biology and evolution*, 27(8):1955–1962.

Wickham, H. and Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527):355.

Figures and tables

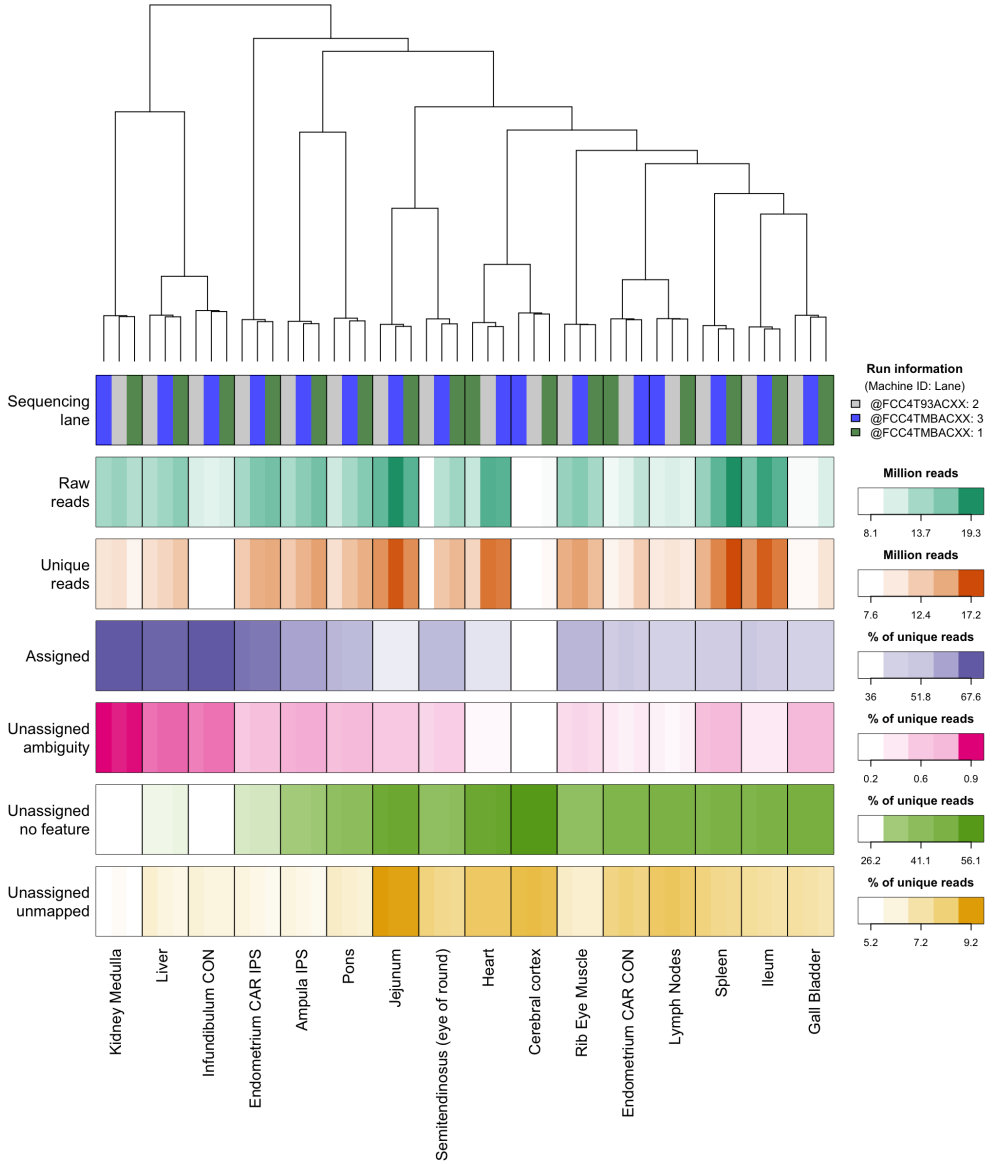


Figure 1. Hierarchical clustering of each individual sequencing run and associated sequencing/mapping statistics. CON stands for 'contralateral to the corpeus Luteum', CAR stands for 'caruncular regions' and IPS stands for 'ipsilateral to the corpeus Luteum'.

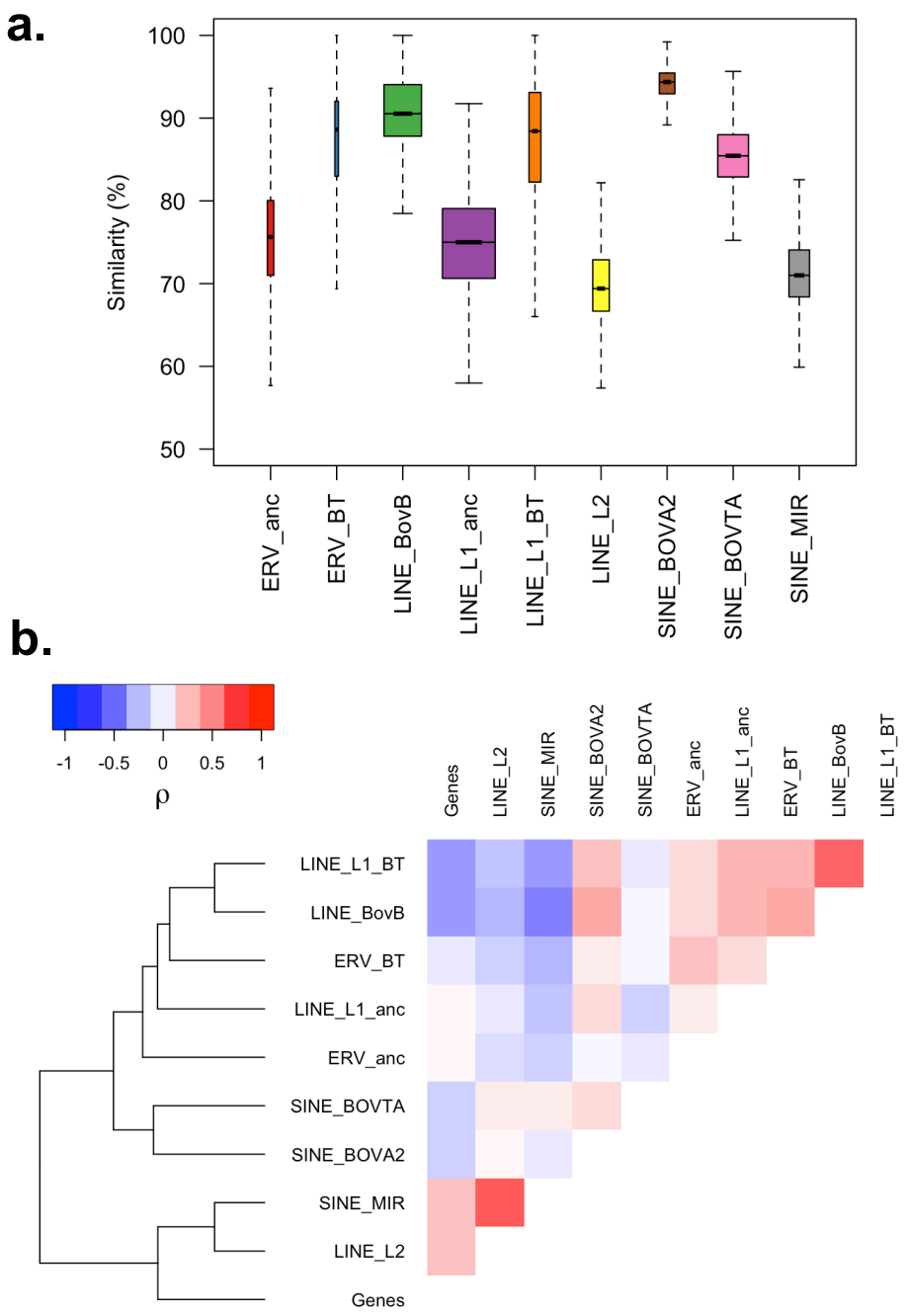


Figure 2. The bovine TE landscape. **a.** The age distribution of of our identified TE groups. The y-axis represents % similarity to consensus sequence and is a rough indicator of TE age, where youger TEs show higher levels of similarity to their consensus sequences. Box width is directly proportional to the total number of instances for each TE group. **b,**) Pairwise correlation analysis between each of our TE groups.

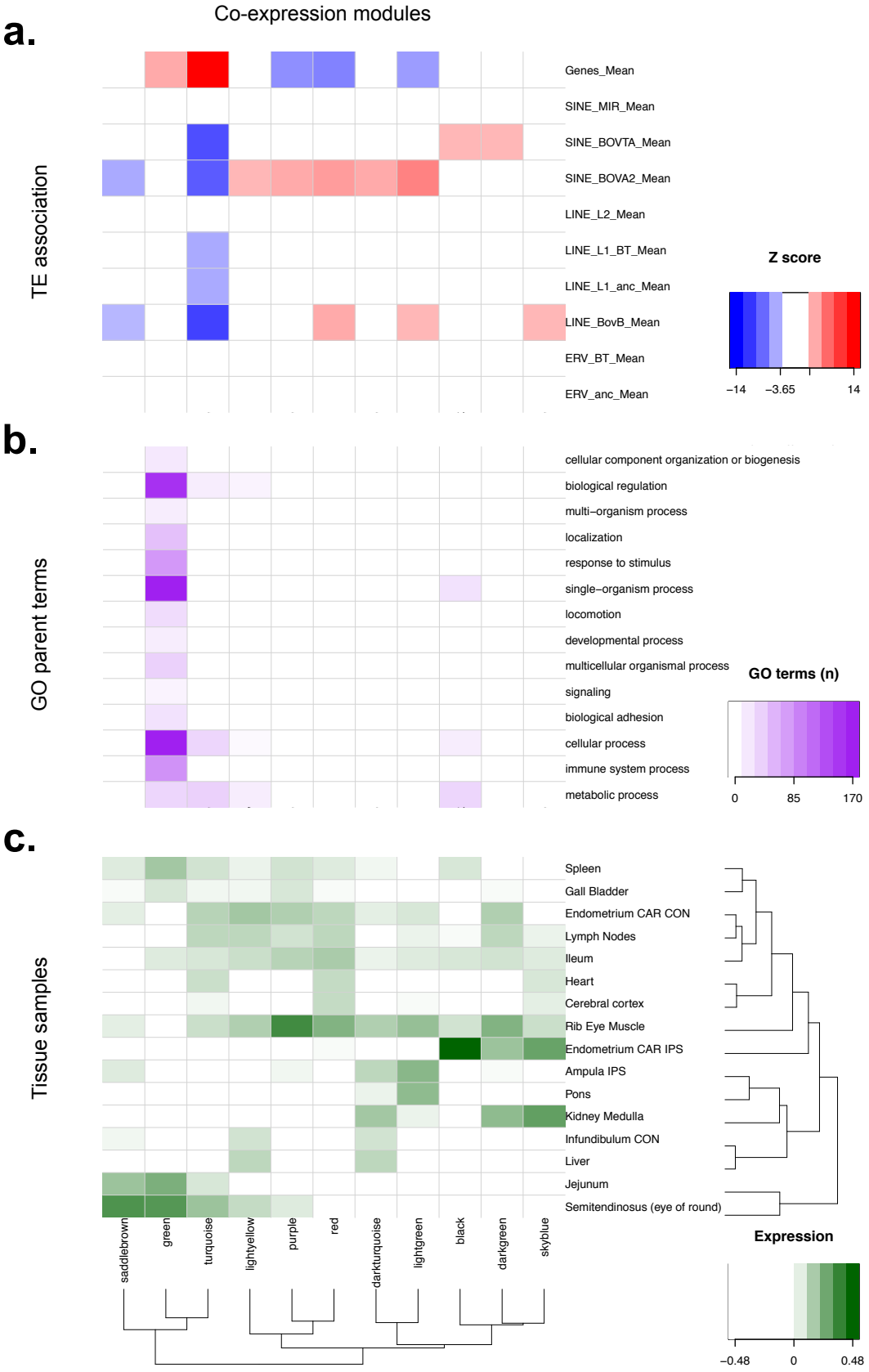

a.
Co-expression modules
TE association
Genes_Mean
SINE_MIR_Mean
SINE_BOVTA_Mean
SINE_BOVA2_Mean
LINE_L2_Mean
LINE_L1_BT_Mean
LINE_L1_anc_Mean
LINE_BovB_Mean
ERV_BT_Mean
ERV_anc_Mean
Z score
−14
−3.65
14
b.
GO parent terms
cellular component organization or biogenesis
biological regulation
multi−organism process
localization
response to stimulus
single−organism process
locomotion
developmental process
multicellular organismal process
signaling
biological adhesion
cellular process
immune system process
metabolic process
GO terms (n)
0
85
170
c.
Tissue samples
Spleen
Gall Bladder
Endometrium CAR CON
Lymph Nodes
Ileum
Heart
Cerebral cortex
Rib Eye Muscle
Endometrium CAR IPS
Ampula IPS
Pons
Kidney Medulla
Infundibulum CON
Liver
Jejunum
Semitendinosus (eye of round)
saddlebrown
green
turquoise
lightyellow
purple
red
darkturquoise
lightgreen
black
darkgreen
skyblue
Expression
−0.48
0
0.48

Figure 3. TE association with co-expression models, their associated biological processes and their tissue-specific expression patterns. **a,** Z score for regional TE association of each co-expression module's member genes. Z scores were generated using a permutation approach consisting of 10,000 iterations (methods). Z scores are shown if they are > 3.65 standard deviations from the mean, based on correction for multiple testing at a FDR < 0.05. **b,** The number of significantly enriched child GO terms that belong to top level BP parent terms. **c,** Eigengene expression of co-expression modules across each of our samples.

Chapter 5

Bovine *NK-lysin*: Copy number variation and functional diversification

Retrotransposon accumulation in itself has the ability to cause large structural rearrangements. The bovine NK-lysin gene was found to be incorrectly assembled in the bovine reference. In fact, there were three additional copies of the NK-lysin gene that were specific to the bovine lineage. For each tandem duplicate copy, I characterised the retrotransposons surrounding their breakpoints and found strong enrichment for bovine-specific SINEs. This was consistent with tandem segmental duplication caused by repeat mediated non-homologous recombination.

Statement of Authorship

Title of Paper	Bovine NK-lysin: Copy number variation and functional diversification
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Chen, Junfeng, John Huddleston, Reuben M. Buckley, Maika Malig, Sara D. Lawhon, Loren C. Skow, Mi Ok Lee, Evan E. Eichler, Leif Andersson, and James E. Womack. "Bovine NK-lysin: Copy number variation and functional diversification." *Proceedings of the National Academy of Sciences* 112, no. 52 (2015): E7223-E7229.

Co-Author

Name of Co-Author (Candidate)	Reuben Buckley
Contribution to the Paper	Annotated repetitive elements, analysed their genomic distributions, provided figures for the manuscript.
Overall percentage (%)	10%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the third author of this paper.
Signature	Date 13/4/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the c[illegible] he thesis, and

iii. the sum of all co-author contrit[illegible] idate's stated contribution

Name of Co-Author	LOREN Skow
Contribution to the Paper	I participated in the research described in this paper by providing guidance in experiments and data collection, helped critique experimental designs, review data, and [illegible] author on this paper.
Signature	Date 17 April 2017

Name of Co-Author	Mi Ok Lee
Contribution to the Paper	I provided expertise for RNA expression analysis.
Signature	Date 4/11/2017

Name of Co-Author	vhon		
Contribution to the Paper	Provided expertise and supervision for bacterial cultures and data analysis and provided editorial review of the resultant manuscript.		
Signature		Date	April 17, 2017

Name of Co-Author	omack		
Contribution to the Paper	Contributed to design of the project and routine analysis of results.		
Signature		Date	4-19-17

Name of Co-Author	Anderson		
Contribution to the Paper	Contributed to design of the project and interpretation of data		
Signature		Date	May 9, 2017

Name of Co-Author	Junfeng Chen		
Contribution to the Paper	I contributed to the design of the project. performed research and wrote the paper.		
Signature		Date	5-09-2017

Name of Co-Author	Evan Eichler		
Contribution to the Paper	Contributed to sequence characterization of duplicate genes.		
Signature		Date	6/9/17

Name of Co-Author	John Huddleston		
Contribution to the Paper	sequence characterization of duplicate genes		
Signature		Date	6/20/2017

Name of Co-Author	Maika Malig
Contribution to the Paper	Isolated BAC clones, created sequenced libraries and sequenced libraries in ES II machine
Signature	Date 6/30/17

Bovine *NK-lysin*: Copy number variation and functional diversification

Junfeng Chen[a], John Huddleston[b,c], Reuben M. Buckley[d], Maika Malig[b], Sara D. Lawhon[a], Loren C. Skow[e], Mi Ok Lee[a], Evan E. Eichler[b,c], Leif Andersson[e,f,g], and James E. Womack[a,1]

[a]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843; [b]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [c]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195; [d]School of Biological Sciences, University of Adelaide, Adelaide 5005, Australia; [e]Department of Veterinary Integrative Biosciences, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843; [f]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, SE 75123, Sweden; and [g]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, SE 75007, Sweden

Contributed by James E. Womack, November 20, 2015 (sent for review November 5, 2015; reviewed by Denis M. Larkin and Harris A. Lewin)

NK-lysin is an antimicrobial peptide and effector protein in the host innate immune system. It is coded by a single gene in humans and most other mammalian species. In this study, we provide evidence for the existence of four *NK-lysin* genes in a repetitive region on cattle chromosome 11. The *NK2A*, *NK2B*, and *NK2C* genes are tandemly arrayed as three copies in ∼30–35-kb segments, located 41.8 kb upstream of *NK1*. All four genes are functional, albeit with differential tissue expression. *NK1*, *NK2A*, and *NK2B* exhibited the highest expression in intestine Peyer's patch, whereas *NK2C* was expressed almost exclusively in lung. The four peptide products were synthesized ex vivo, and their antimicrobial effects against both Gram-positive and Gram-negative bacteria were confirmed with a bacteria-killing assay. Transmission electron microcopy indicated that bovine *NK-lysins* exhibited their antimicrobial activities by lytic action in the cell membranes. In summary, the single *NK-lysin* gene in other mammals has expanded to a four-member gene family by tandem duplications in cattle; all four genes are transcribed, and the synthetic peptides corresponding to the core regions are biologically active and likely contribute to innate immunity in ruminants.

NK-lysin | antimicrobial peptides | gene family expansion | segmental duplication | copy number polymorphism

Antimicrobial peptides (AMPs) are effector molecules in the innate immune system and are widespread in all kingdoms of life (1, 2). Human granulysin (*GNLY*) and pig *NK-lysin* are orthologs and belong to the same group of AMPs (3, 4). They are secreted from the granules of cytotoxic T lymphocytes and natural killer (NK) cells and are active against a wide spectrum of microorganisms including Gram-positive and Gram-negative bacteria, fungi, protozoa, viruses, and even tumor cells (5–11). *NK-lysin* orthologs have been identified and characterized in many species, including human, pig, cattle, horse, water buffalo, and several species of birds (12–15). Bovine *NK-lysin* was first reported a decade ago (16), when two bovine cDNA fragments were obtained from each of four different cows. It was unclear whether the detected sequences, *Bo-lysin 89* and *Bo-lysin 62*, were from two different *NK-lysin* genes or were alleles of a single gene. Also, multiple variants of *NK-lysin* sequences exist in the bovine nucleotide database, suggesting the existence of more than one copy of *NK-lysin* in the cattle genome (Fig. S1 and Table S1).

Copy number variation (CNV) is a common form of structural variation in animal genomes. Several whole-genome CNV analyses have been carried out among different breeds of cattle, and two independent studies suggested that bovine *NK-lysin* is in a CNV region (17, 18). Duplications (>1 kb) that are highly identical (90%) are known as "segmental duplications." Segmental duplications are common in mammalian genomes and are highly copy-number variable, serving as one of the principal mechanisms of gene family expansion (19) which can provide substrates for neofunctionalization and development (20, 21).

Sequencing of the cattle genome (22) revealed that multiple immune-related genes are expanded in copy number in cattle as compared with humans and mice. These include genes coding AMPs such as the cathelicidins and β-defensins, members of the IFN gene family, C-type lysozyme, and lipopolysaccharide-binding protein (*ULBP*) (23–28). Expansion of these gene families potentially can give rise to new functional paralogs with implications in the unique gastric physiology of ruminants or in disease resistance in a herd environment. Here we demonstrate that there are four copies of *NK-lysin* in cattle; three related copies are located in tandem within ∼30–35-kb regions of segmental duplication, whereas the fourth copy is located 41.8 kb downstream. All four genes show tissue-specific expression, and the product of each of the four genes displays antimicrobial activity against both Gram-positive and Gram-negative bacteria by the mechanisms of pore formation and cell lysis.

Results

Analysis of Cattle Homozygous at the *NK-Lysin* Locus. A search of the National Center for Biotechnology Information (NCBI) bovine nucleotide database identified seven different *NK-lysin*–related sequences (Table S1), and a phylogenetic analysis of the sequences showed four clades that potentially represented four different bovine *NK-lysin* genes. We designated these genes *NK1*, *NK2A*, *NK2B*, and *NK2C* (Fig. S1). *NK2A*, *NK2B*, and *NK2C* were closely related to each other and were divergent from *NK1*. The genes corresponding to *NK1* and *NK2A* have been annotated previously as *uncharacterized LOC616323* (gene ID:

Significance

The cattle genome contains expanded families of several genes involved in innate immunity. A single copy of the *NK-lysin* gene is annotated in the genomes of most mammals, including humans, but this study identified a family of *NK-lysin* genes in cattle consisting of four functional members. Although this family mirrors the numerical expansion of other immune-related genes, including interferons, defensins, and cathelicidins, in the cattle genome, we also see a diversification of function exhibited by differential tissue expression in the gene family. The current state of this site in the bovine genome appears to capture the evolutionary transition from copy number variation to the fixation of novel gene function within a segmentally duplicated region.

Author contributions: J.C., E.E.E., L.A., and J.E.W. designed research; J.C., J.H., R.M.B., and M.M. performed research; S.D.L. and L.C.S. contributed new reagents/analytic tools; J.C. and M.O.L. analyzed data; and J.C. wrote the paper.

Reviewers: D.M.L., Royal Veterinary College; and H.A.L., University of California, Davis.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this article has been deposited in the National Center for Biotechnology Information database (accession no. KT715031).

[1]To whom correspondence should be addressed. Email: jwomack@cvm.tamu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1519374113/-/DCSupplemental.

LOC616323) and *Bovine GNLY* (gene ID: 404173), respectively, in the bovine reference genome assembly UMD 3.1 of the University of California, Santa Cruz genome browser. These two genes are tandemly arranged on chromosome 11, whereas *NK2B* and *NK2C* are absent in the current genome assemblies. To confirm the authenticity of the *NK2A*, *NK2B*, and *NK2C* sequences, we designed a pair of primers (Bo-lysin F: Bo-lysin R) from the conserved region of these genes. To minimize the effects of allelic variation in the analysis, we selected four Holstein cattle homozygous for this region based on genome-wide association study genotyping results with the 770K HD SNP array (29). The SNP array contained 29 SNPs between the two genes flanking the *NK-lysin* region, *ATOH8* (gene ID: 616225) and *SFTPB* (gene ID: 507398). The PLINK program was used to identify individuals that were homozygous at all 29 SNP sites, and four cattle (2527, 2796, 2822, and 3850) with different haplotypes were selected for further analysis. The number of the sequenced clones and the different sequences achieved from each individual are listed in Table S2. In total, five different sequences (Seq1–5) were recovered from these four individuals. The five sequences formed three clades, corresponding to the *NK2A*, *NK2B*, and *NK2C* genes, and were divergent from *NK1* (Fig. 1). Three different arrangements of *NK-lysin* genes were observed in this study. Two sequences from the *NK2A* cluster were detected in individual 2527. If the individual 2527 was homozygous across the *NK-lysin* region, at least two copies of *NK2A* were present in this animal. Despite the large number of clones sequenced from both individuals 2822 and 3850, we found no *NK2B*-related clones, and we could not obtain *NK2B* amplicons with the *NK2B*-specific primer, suggesting the absence of the *NK2B* gene in these animals.

BAC Clone Sequencing Identified Four *NK-Lysin* Genes. The precise number of genes in the bovine *NK-lysin* family and their genomic organization were determined by sequencing two overlapping BAC clones covering the *NK-lysin* region. The clones were isolated from the CHORI-240 Bovine BAC Library and were sequenced with P4/C2 chemistry on the PacBio RS. Despite a sequencing coverage depth of >700× for both BACs after the first round of sequencing, each BAC was assembled into six contigs because of the presence of highly repetitive sequences. After a second round of sequencing, the average coverage was increased to ~1,310–1,551×; however, three contigs were still generated from CH240-372P1, and two contigs were generated from CH240-27G22. Because these two BAC clones overlap, we were able to perform a final de novo assembly of all sequencing data. This analysis produced a two-contig assembly in which the two contigs overlapped by ~2 kb at 100% identity. These two contigs subsequently were joined into a single contig, resulting in a linear supercontig of 227,063 bp covering the whole bovine *NK-lysin* region. Overall, the assembled contig (Bo-NK) was longer than the current genome assembly by ~38 kb, where the corresponding reference sequence was 189,124 bp (Bos_taurus_UMD_3.1 Chr. 11: 48,986,139–49,175,262 bp). The difference in length was caused primarily by misassemblies in the reference genome, in which repetitive regions containing the *NK2B* and *NK2C* genes were collapsed (Fig. 2*A*).

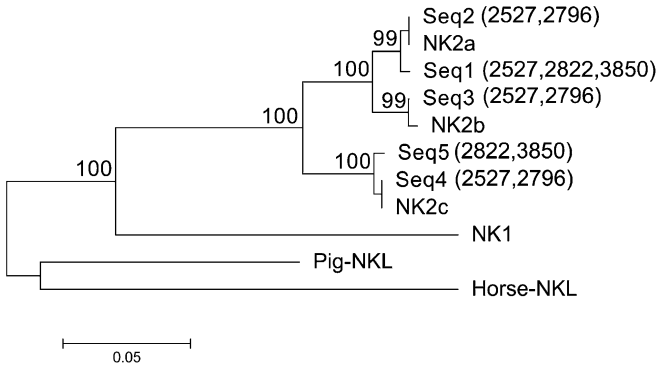
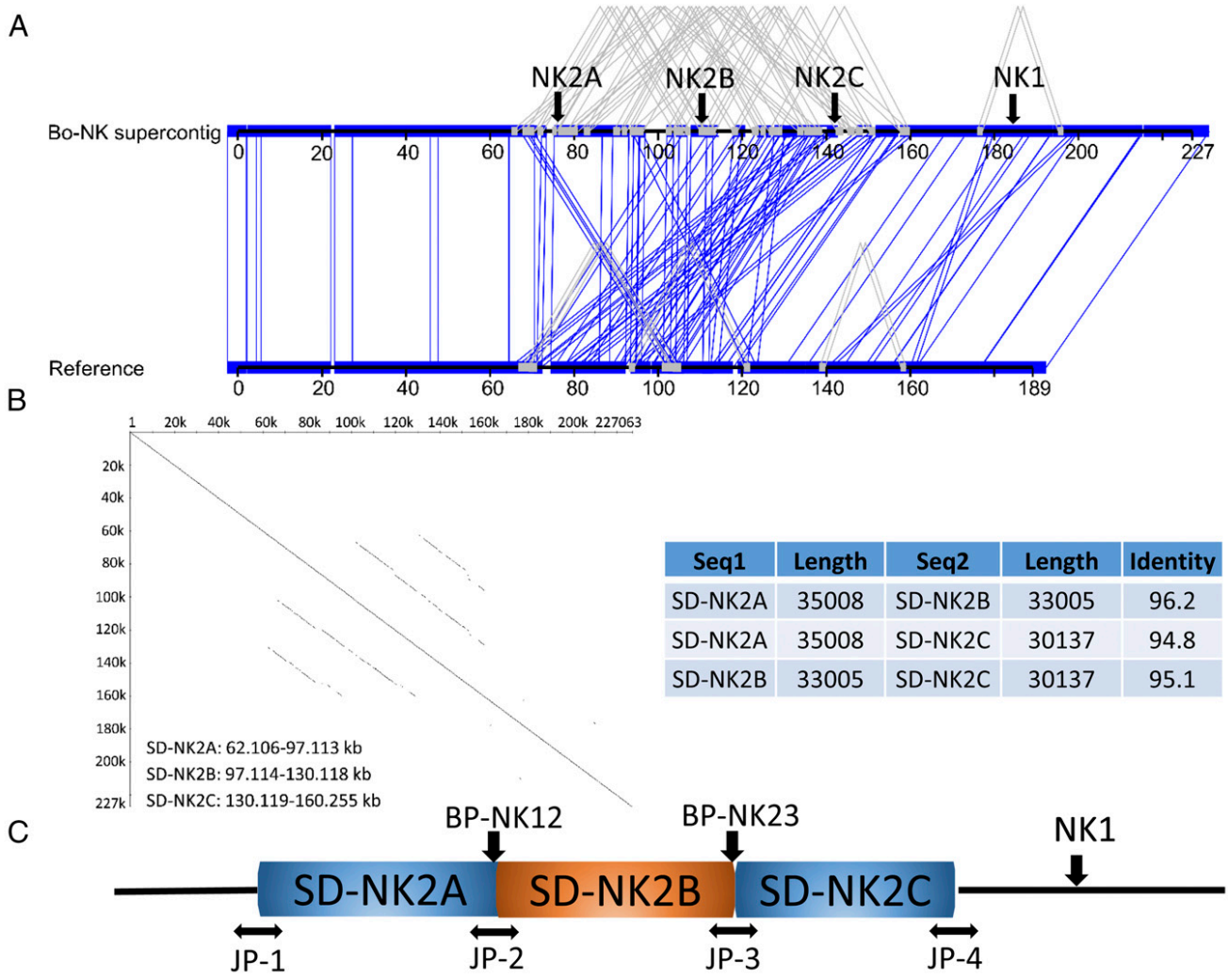


Fig. 1. *NK2A, NK2B,* and *NK2C* nucleotide sequence analysis in four homozygous individuals (2527, 2796, 2822, and 3850). Five different clone sequences (Seq. **1**–5) from four individuals were phylogenetically analyzed with four bovine *NK-lysin* reference sequences (*NK1*, *NK2A, NK2B,* and *NK2C*) and corresponding pig (Pig-NKL) and horse (Horse-NKL) orthologs by the MEGA 6.0. Bootstrap values are shown at branch points.

Dot plot analysis of the Bo-NK contig against itself revealed three segmental duplications with ~95% sequence identity (SD-NK2A: 62.1–97.1 kb; SD-NK2B: 97.1–130.1 kb; and SD-NK2C: 130.1–160.3 kb), each containing one *NK-lysin* gene; *NK1* was 41.8 kb downstream from the *NK2C* gene (Fig. 2*B*). Because the SD-NK2C lacked the right end of the duplicated fragment and was shorter than SD-NK2A and SD-NK2B, the flanking sequence of junction point 4 (JP-4) was different from the other three breakpoints (JP-1, JP-2, and JP-3) (Fig. 2*C*). To confirm the accuracy of the Bo-NK contig, we tested four primer pairs at each junction point using genomic DNA of L1 Domino 99375 (donor for the CHORI-240 Bovine BAC Library). Sanger sequencing showed that JP-1, JP-3, and JP-4 PCR products were perfectly aligned with the Bo-NK contig, but there were six mismatches out of 567 nucleotides between the JP-2 PCR product and the Bo-NK contig. The amplicon of another primer pair (BP-1) was sequenced by Sanger to determine whether these six mismatches were the result of an error in the PacBio sequencing. Sanger sequencing verified six sequencing errors at the BP-NK12 breakpoint in the Bo-NK contig. The Bo-NK contig therefore represented the correct assembly of the bovine *NK-lysin* region and demonstrated that four *NK-lysin* genes are located in this region on cattle chromosome 11. Complete genomic sequences of four *NK-lysin* genes were compared with determine genetic organization and structure (Fig. S2). All four bovine *NK-lysin* genes contain five exons, as is consistent with the architecture of human and pig orthologs. The exon sizes were comparable among the four genes, but the introns of *NK1* were larger than the introns from the other genes, accounting for the larger genomic size of *NK1* (Fig. S2*A*). *NK2A*, *NK2B*, and *NK2C* are about 95% identical to each other but are only 85% identical to *NK1*. The predicted amino acid compositions of the four bovine *NK-lysin*s show high sequence identity and include six cysteine residues, which are conserved among *NK-lysin* molecules in other animals (Fig. S2*B*). Phylogenetic analysis of the full coding sequences of the four bovine *NK-lysins* with *NK-lysin* orthologs in humans, pig, horse, sheep, and goat revealed that the expansion of the *NK-lysin* gene family is seen only in the ruminants, suggesting the divergence of the *NK1* and *NK2* cluster in the ancestor of cattle, sheep, and goats (Fig. S3).

Analysis of Repetitive Sequences Within the Bovine NK-Lysin Gene Family. Repetitive sequences usually are associated with recombination hotspots in the human genome (30), and chromosomal instability caused by mispairing between such repeats at breakpoints is responsible for several diseases (31, 32). To gain more insight into the mechanism of *NK-lysin* expansion in cattle, we analyzed the distribution of repeat elements within this region. The distributions of different repeat classes within the assembled contig are shown in Fig. 3*A* and are summarized in Table S3. Overall, the downstream region of each breakpoint is more repetitive than the upstream region, and the flanking sequences of *NK1* are highly repetitive, consisting of a large percentage of long, interspersed nuclear elements (LINES), which is distinct from the rest of the region within this gene family. Several repeat families are overrepresented within the *NK-lysin*



Seq1	Length	Seq2	Length	Identity
SD-NK2A	35008	SD-NK2B	33005	96.2
SD-NK2A	35008	SD-NK2C	30137	94.8
SD-NK2B	33005	SD-NK2C	30137	95.1

```
JP-2   GAGA--GCTT GGGTTTGATT CCTGGACAGG GAACTAAGAT CCCACAGGC- TGTGGGGCAA
JP-3   ....--.... .......... .......... .......... .........- ..........
JP-1   ACT.CTA... CCAACAC.GC GG.T.C.... .......... ......T..- ....T.....
JP-4   TG.CAT...- ..AGC.CTGC TGCT.CT--. CTG.....T. G.TTT..T.A ...CC.A.TC
```

Fig. 2. BAC clone analysis by PacBio sequencing. (*A*) Sequence comparison between the Bo-NK supercontig and the genome assembly (Bos_taurus_UMD_3.1.1). Mismatches (vertical blue lines), internal duplications (gray boxes), and four *NK-lysin* gene loci (arrows) are indicated. (*B*) Dot plot analysis of the Bo-NK supercontig against itself. (*C*) Genomic organization of the bovine *NK-lysin* gene family and identified breakpoints (BP). The flanking sequence of JP-2 was used as the reference sequence.

region, including two ancient mammalian L1 families, two LTR families, and four ruminant/bovine-specific short, interspersed nuclear element (SINE) families (BOVTA, BTALUL2, CHR-2_BT, and CHR-2A) (Fig. S4). Because of the enrichment of SINEs around junction points, we plotted the distribution of several ruminant/bovine-specific repeat families within 5 kb upstream and downstream of each junction point (Fig. 3*B*). The adjacent downstream regions of JP-1, JP-2, and JP-3 are enriched with SINES, especially the BOVTA element. BOVTA elements form a bovine-specific repeat family analogous to the primate ALU repeat family, which usually is associated with segmental duplications in humans (33). These results demonstrate that the fragments flanking breakpoints share high homology and could contribute to unequal crossover during meiosis and structural instability within the bovine *NK-lysin* gene family.

Tissue Expression of the Bovine *NK-Lysin* Genes. To test whether all the identified bovine *NK-lysin* genes are expressed and display the same expression profile, we compared the mRNA levels of each gene among five tissues, including lung, thymus, spleen, respiratory lymph node (RLN), and intestine Peyer's patch (IPP). Real-time PCR analysis demonstrated that all four bovine *NK-lysin* genes are expressed, but each exhibits a tissue-specific expression profile (Fig. 4). *NK1* and *NK2A* genes are highly expressed in the IPP but are expressed at extremely low levels in the lung. The difference was greater than 100-fold. *NK2B* is more generally expressed, with highest levels in the IPP and lung. A distinct expression pattern was observed for *NK2C*, which was expressed at highest level in the lung, indicating a potential novel function.

Antimicrobial Effects of Bovine *NK-Lysin* Peptides. Antimicrobial capacities of synthetic forms of four bovine *NK-lysin* peptides

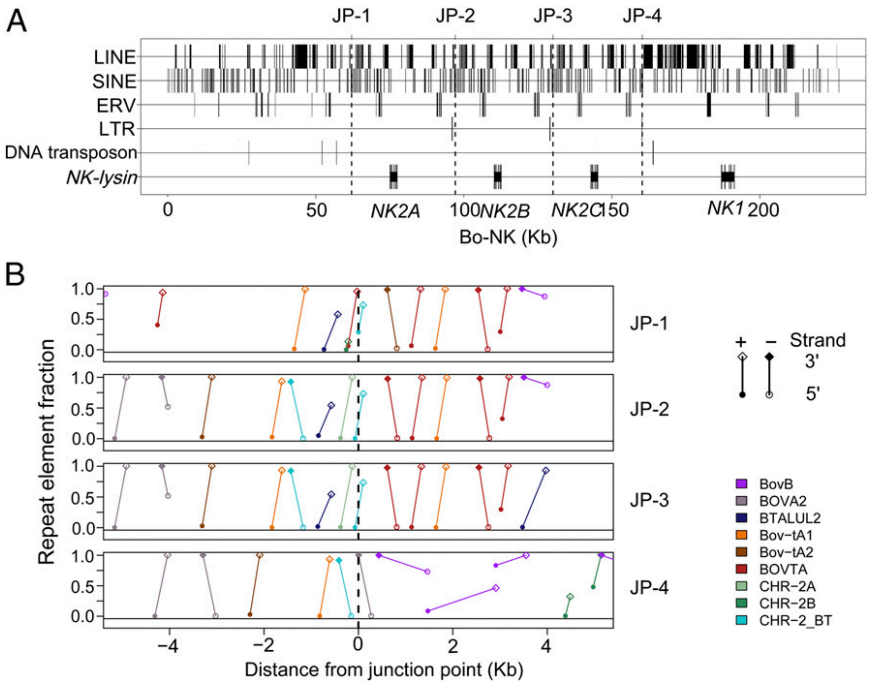


Fig. 3. Repeat element analysis. (*A*) Distribution of repeat classes within the assembled supercontig Bo-NK. Four junction points and genes are indicated. (*B*) Distribution of SINEs within 5 kb upstream and downstream of each junction point. The portion of each element relative to its consensus sequence is shown on the *y* axis.

were tested against both the Gram-positive bacteria *Staphylococcus aureus* and the Gram-negative bacteria *Escherichia coli*. All peptides were effective against both bacterial strains at nanomolar concentrations, although Gram-negative *E. coli* was more susceptible (Fig. 5*A*). At the lowest concentration of 0.05 μm, an ∼10-fold decrease in viable *E. coli* cells was observed, and bacterial numbers were reduced from initial 10^6 cfu/mL to less than 10^4 cfu/mL after incubation with 1 μm of *NK2A*, *NK2B*, or *NK2C* molecules for 2 h. Even fewer cells (400 cfu/mL) survived incubation with 1 μm *NK1* peptide. All peptides were less active against the Gram-positive *S. aureus*. Bacterial numbers were not reduced significantly when incubated at peptide concentrations up to 0.1 μm for any of the four peptides. At 0.5 μm, all peptides produced ∼10-fold cell loss. At the concentration of 1 μm, the *NK1* molecule was stronger than the other three peptides and reduced *S. aureus* numbers by ∼100-fold. Although the peptides differed in their ability to kill Gram-positive and Gram-negative bacterial strains, the *NK1* peptide showed strongest antimicrobial effects against both strains. Despite having a less-positive charge, the *NK1* peptide is more hydrophobic than the other three peptides (Table S4), perhaps explaining its stronger antimicrobial effects against the tested bacterial strains.

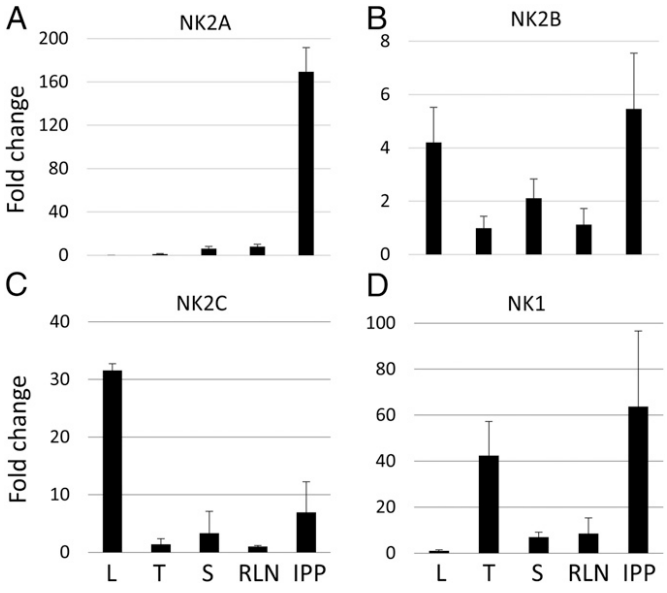


Fig. 4. Expression of four bovine *NK-lysins* in lung (L), thymus (T), spleen (S), RLN, and IPP. The expression of each gene in the tissue that exhibited the lowest expression level was assumed to be 1. The average expression levels and SDs were calculated from three healthy individuals. (*A*) *NK2A*. (*B*) *NK2B*. (*C*) *NK2C*. (*D*) *NK1*.

The effects of bovine *NK-lysin* molecules on the *E. coli* cell membrane were investigated by transmission electron microscopy (TEM) (Fig. 5*B*). Specifically, the membrane integrity and intracellular structure of untreated *E. coli* cells and cells treated with 5 μm *NK1* peptide were compared and analyzed. The observed differences in the membrane ultrastructure caused by treatment with *NK1* peptide were obvious. Most of the untreated cells maintained a normal cell shape with an intact cytoplasmic membrane and full cytoplasmic contents (Fig. 5 *B*, *a*), whereas treated cells had characteristic expansion of the periplasmic space with shrinkage of the cytoplasmic compartment (Fig. 5 *B*, *b*). The cytoplasm of treated cells was less electron dense, with clear zones, indicating the disruption of the cell membranes and leakage of intracellular contents. Protruding bubbles were observed from the membrane of treated cells (Fig. 5 *B*, *d*) whereas outer membranes of untreated cells displayed a uniform appearance with slightly waved membranes (Fig. 5 *B*, *c*). Statistical analysis confirmed that the average electron density of untreated cells was significantly ($P < 0.001$) stronger than the treated ones (Fig. 5 *B*, *e*). The results from this assay demonstrated the lytic action of bovine *NK-lysin* peptides, which may directly cause pore formation in the cell membrane.

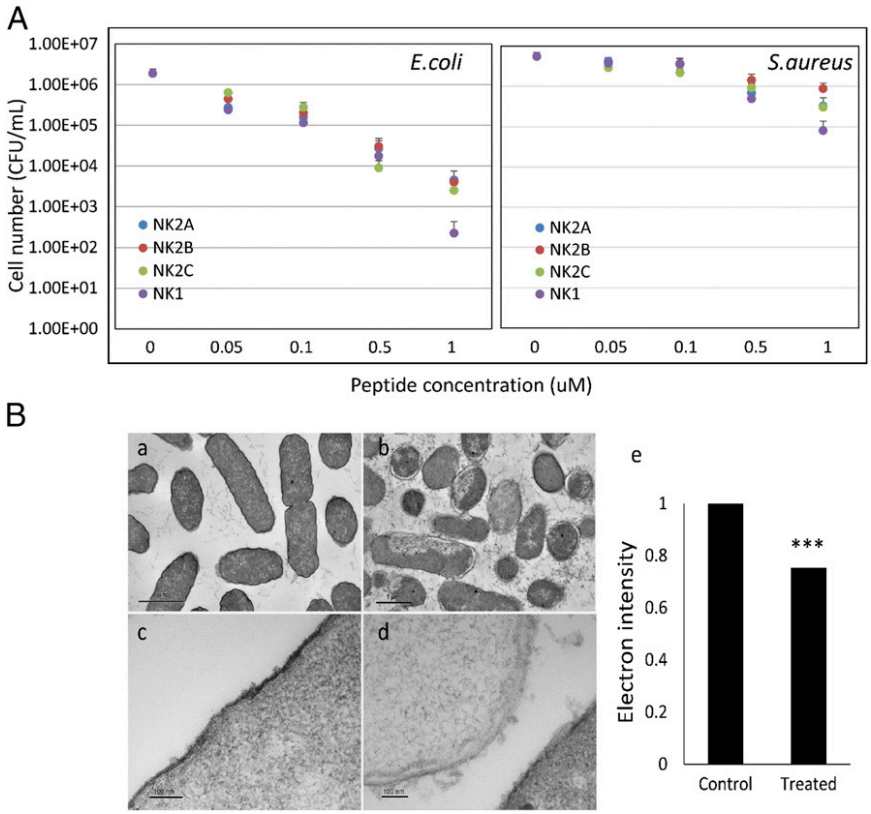


Fig. 5. (*A*) Antimicrobial activities of four bovine *NK-lysin* peptides against Gram-negative *E. coli* and Gram-positive *S. aureus*. Cell viability was analyzed by comparing the surviving cells after peptide treatment with the control cells. Error bars represented the SDs calculated from four biological replications. (*B*) Transmission electron micrographs of *E. coli* cells with and without 5 μM *NK1* peptide treatment. (*a* and *c*) Control cells. (*b* and *d*) Cells treated with 5 μM *NK1* peptide for 20 min. (e) Comparison of the average electron intensity of 30 cells in the control and *NK1*-treated cell groups.

Discussion

In this study, we provide evidence for tandem duplications of three *NK-lysin* genes, likely derived from an ancestral fourth copy located ~41.8 kb downstream on cattle chromosome 11. Conserved features of *NK-lysin* orthologs, including the presence of five exons/four introns, six well-conserved cysteine residues, and a high proportion of positively charged amino acids, exist in all four bovine *NK-lysin* genes. The genome context flanking the bovine *NK-lysin* gene family demonstrated conserved syntenies with the *granulysin* region of human and most other mammalian genomes. The human *granulysin* gene maps to chromosome 2 centromeric to *SFTPB* (surfactant protein B) and *USP39* (ubiquitin-specific peptidase 39) and telomeric to *ATOH8* (atonal homolog 8) and *ST3GAL5* (ST3 β-galactoside α-2,3-sialyltransferase 5). Similarly, the bovine *NK-lysin* gene family maps centromeric to *SFTPB* and *USP39* and telomeric to *ATOH8* and *ST3GAL5* on chromosome 11. The conserved genome context implies that no major interchromosomal genomic reorganization has occurred in this region since the divergence of the ancestors of cattle and humans.

The arrangement of *NK2A*, *NK2B*, and *NK2C* as head-to-tail tandem triplicates is consistent with the predominate duplication pattern observed in cattle and other mammals including mouse, rat, and dog and is in contrast to the archetypical organization of interspersed duplications in higher primates (34–39). Segmental duplication with subsequent differentiation is the major mechanism of gene family expansion. Acting as the substrates of genome evolution, regions of segmental duplication also are particularly unstable and are hotspots of CNV (37, 38, 40–42). Our analysis of homozygous Holstein cattle revealed copy number polymorphism of *NK2B* and potential copy number polymorphism of *NK2A* in contrast to the BAC sequence contributed by a Hereford bull. We then investigated the features of sequences flanking each breakpoint and found that the fragments downstream of each breakpoint were highly repetitive. These highly repetitive regions share high sequence homology and potentially drive rearrangements among the genetic elements flanked by these repeats; these rearrangements can result in deletions or duplications of genomic fragments. Therefore further studies are suggested to investigate the extent of CNV within and between breeds of cattle in all four bovine *NK-lysins* and haplotype structures within this gene family.

In contrast to the single copy of *NK-lysin* gene in most species including human, pig, chicken, and horse, four *NK-lysin* genes cluster in a region with highly repetitive sequences in the cattle genome. To our knowledge, cattle are the first mammals in which multiple *NK-lysin* genes have been found, and this observation is consistent with the gene family expansions in cattle for several other genes related to innate host immunity, such as the defensins, cathelicidins, and interferons (23–25, 27). Perhaps reflecting an evolutionary strategy to deal with the substantial number of pathogens and the increased risk of infections in the rumen of cattle, the enlarged gene families encoding the AMPs may be selected to meet an increased demand (22). It has been reported that some duplicates of an immunity-related gene exhibit nonimmune functions in cattle, such as the roles of the lysozyme genes in both the immune and digestive systems (22). Although *NK-lysin* orthologs are predominately expressed in the IPP in most species, the bovine *NK2C* gene is expressed at the highest level in lung, implying a potential novel function in the bovine respiratory system.

Bacteria-killing assays revealed that the synthetic peptides from the functional regions of four bovine *NK-lysin* genes are active against both Gram-positive and Gram-negative bacterial strains at

the very low concentration of 0.05 μm. Therefore we provided four potential candidate templates for the development of new antibacterial drugs. However, the size of a peptide is of utmost importance in determining whether it is a feasible antimicrobial drug, and the bovine *NK-lysin* molecules in this study covered the whole functional region of helices 2 and 3 in the genes, which consisted of 30 residues. Further studies are necessary to determine the activities of shortened bovine *NK-lysin* peptides.

Materials and Methods

Analysis of Homozygous Animals. All identified *NK-lysin*-related sequences from the NCBI bovine nucleotide database were subjected to phylogenetic analysis by ClustalW. Primer 3 was used to design a pair of primers (Bo-lysin) within the conserved region of the *NK2A*, *NK2B*, and *NK2C* clusters (Table S5). Four Holstein cattle which were homozygous at all SNP sites across the entire *NK-lysin* region, based on genotyping with the bovine 770K HD SNP array (29), were used in this analysis. The Bo-lysin amplicons from each of the four homozygotes were cloned into the pCR4 Blunt-TOPO vector (Life Technologies) for sequencing (Beckman Coulter Genomics). Only sequences present at least three times among the clones from a single individual were used for analysis. All sequences were analyzed phylogenetically with the corresponding reference sequences of *NK1* and *NK2A–C* by MEGA 6.0 (43); pig and horse *NK-lysin* sequences were included as outgroups. The absence of *NK2B* in individuals 2822 and 3850 were confirmed further by PCR with *NK2B*-specific primers (Gs-NK2B).

BAC Clone Sequencing. Two overlapping BAC clones were selected from the CHORI-240 Bovine BAC Library, and confirmation of *NK-lysin* inclusion was conducted with the Bo-lysin primers. BAC sequencing was carried out with single-molecule real-time (SMRT) sequencing technology (Pacific Biosciences), as described previously (44). Each clone was sequenced twice in two separate SMRT cells. De novo assembly of the data from each SMRT cell and from the combined two SMRT cells from each clone was performed following the standard SMRT Analysis (v. 2.0.1) pipeline. A further de novo assembly was attempted using combined data from all four SMRT cells, and the final contigs were joined into a single supercontig using Sequencher (Gene Codes Corporation). The supercontig then was compared with the reference sequence using the miropeats alignment in Parasight (45), and further dot plot analysis of the supercontig was implemented by UniproUGENE (46, 47). Four pairs of primers specific for each putative junction point (JP-1, JP-2, JP-3, and JP-4) were tested in the genomic DNA of L1 Domino 99375 to validate the BAC assembly.

Repeat Element Analysis. Repeat elements within the Bo-NK supercontig and the UMD_3.1.1 assembly were identified and annotated using CENSOR with a bovine-specific library downloaded from Repbase that included ancestral sequences (48, 49). To estimate the density of each repeat family within the whole-genome assembly (UMD_3.1.1), the assembled chromosomes were broken into different bins of the same size as the Bo-NK contig (~227 kb), and those consisting of >10% gaps were excluded from the analysis. Repeat density for each repeat family with more than five copies in a bin was represented by the repeat coverage per 1,000 bp. Ambiguous repeat elements at boundaries were assigned to bins based on a minimum 50% repeat length overlap threshold. Overlaps between repeats and bins were identified using the GenomicRanges package from Bioconductor (50, 51). Repeat densities across all bins were used to estimate the empirical cumulative distribution function of each repeat family using the "ecdf" command in R and Bioconductor (52), which then was used to estimate the probability of sampling a bin with a repeat density greater than the repeat density of the Bo-NK supercontig [$P(X > x)$]. A repeat family was overrepresented in the Bo-NK supercontig if $P(X > x)$ was <0.05. Finally, repeat annotation plots were generated using the base graphics system in R (52).

Expression Profiles. Total RNA was extracted from the IPP, lung, thymus, spleen, and RLN of three mixed-breed cattle using the RNeasy Mini kit (Qiagen). RNA then was reverse transcribed into cDNA with a SuperScript II Reverse Transcriptase kit (Invitrogen). Specific Taqman-MGB probes and primers for each gene were designed using Primer Express v.2 (Applied Biosystems) and Primer3. Quantitative PCR was performed in triplicate reactions. The mean threshold cycle value (Ct) of each sample was normalized to the internal control, GAPDH, and the expression profile for each gene was obtained by comparing its normalized Ct value with the calibrator sample in which the gene exhibited the lowest expression level.

Bacteria-Killing Assay. Overnight cultures of Gram-positive *S. aureus* (ATCC 25923) and Gram-negative *E. coli* (ATCC 25922) grown in lysogeny broth (LB) at 37 °C with aeration were subcultured to fresh LB at a ratio of 1:50 and were grown at 37 °C with aeration for another 2.5 h to midexponential phase, washed, and resuspended in potassium phosphate buffer (10 mM, pH 7.4) to a concentration of $3*10^6$ cfu/mL. An aliquot of 110 μL of prepared bacterial cells was incubated with 10 μL buffer or buffer plus peptides at working concentrations of 0.05, 0.1, 0.5, and 1 μM at 37 °C for 2 h and then was plated onto LB agar plates. Colonies of the surviving bacteria were counted manually after overnight incubation at 37 °C.

TEM. One hundred ten microliters of *E. coli* cells (ATCC 25922) ($3 * 10^8$ cfu/mL) were incubated with 10 μL buffer or 5 μM *NK1* peptide at 37 °C for 20 min. Cells were fixed with equal volume of 2.5% glutaraldehyde at room temperature for 2 h and then were washed and placed in 0.1 M sodium cacodylate buffer. The fixed cells were postfixed in 1% OsO4 with 1% $K_4[Fe(CN)_6]$ for 1 h at 4 °C, rinsed with 0.1 M sodium cacodylate buffer followed by dehydration in an ascending ethanol gradient (50, 70, 80, 90, 95, and 100%), and embedded in epoxy resin. Ultrathin sections were obtained with a Leica EM UC6 Ultramicrotome, were poststained with uranyl acetate and lead citrate, and were examined with a Morgagni 268 transmission electron microscope (FEI). Additional image analyses were performed with ImageJ (53). Statistical analysis of the mean electron intensities of 30 cells from both the control and *NK1*-treated groups was performed with Student *t*-test (paired, two-tailed, unequal variances).

ACKNOWLEDGMENTS. We thank Harold Payne for providing assistance and advice on TEM analysis and David L. Adelson for his suggestions on repeat element analysis. This research was supported by Agriculture and Food Research Initiative Competitive Grant 2011-68004-30367 from the US Department of Agriculture National Institute of Food and Agriculture.

1. Zasloff M (2002) Antimicrobial peptides of multicellular organisms. *Nature* 415(6870): 389–395.
2. Hancock RE, Diamond G (2000) The role of cationic antimicrobial peptides in innate host defences. *Trends Microbiol* 8(9):402–410.
3. Peña SV, Hanson DA, Carr BA, Goralski TJ, Krensky AM (1997) Processing, subcellular localization, and function of 519 (granulysin), a human late T cell activation molecule with homology to small, lytic, granule proteins. *J Immunol* 158(6):2680–2688.
4. Andersson M, et al. (1995) NK-lysin, a novel effector peptide of cytotoxic T and NK cells. Structure and cDNA cloning of the porcine form, induction by interleukin 2, antibacterial and antitumour activity. *EMBO J* 14(8):1615–1625.
5. Stenger S, et al. (1998) An antimicrobial activity of cytolytic T cells mediated by granulysin. *Science* 282(5386):121–125.
6. Dieli F, et al. (2001) Granulysin-dependent killing of intracellular and extracellular Mycobacterium tuberculosis by Vgamma9/Vdelta2 T lymphocytes. *J Infect Dis* 184(8):1082–1085.
7. Gansert JL, et al. (2003) Human NKT cells express granulysin and exhibit antimycobacterial activity. *J Immunol* 170(6):3154–3161.
8. Ernst WA, et al. (2000) Granulysin, a T cell product, kills bacteria by altering membrane permeability. *J Immunol* 165(12):7102–7108.
9. Jacobs T, Bruhn H, Gaworski I, Fleischer B, Leippe M (2003) NK-lysin and its shortened analog NK-2 exhibit potent activities against Trypanosoma cruzi. *Antimicrob Agents Chemother* 47(2):607–613.
10. Wang Z, et al. (2000) Bactericidal and tumoricidal activities of synthetic peptides derived from granulysin. *J Immunol* 165(3):1486–1490.
11. Hata A, et al. (2001) Granulysin blocks replication of varicella-zoster virus and triggers apoptosis of infected cells. *Viral Immunol* 14(2):125–133.
12. Hong YH, et al. (2006) Molecular cloning and characterization of chicken NK-lysin. *Vet Immunol Immunopathol* 110(3-4):339–347.
13. Davis EG, Sang Y, Rush B, Zhang G, Blecha F (2005) Molecular cloning and characterization of equine NK-lysin. *Vet Immunol Immunopathol* 105(1-2):163–169.
14. Kandasamy S, Mitra A (2009) Characterization and expression profile of complete functional domain of granulysin/NK-lysin homologue (buffalo-lysin) gene of water buffalo (Bubalus bubalis). *Vet Immunol Immunopathol* 128(4):413–417.
15. Wang Q, Bao B, Wang Y, Peatman E, Liu Z (2006) Characterization of a NK-lysin antimicrobial peptide gene from channel catfish. *Fish Shellfish Immunol* 20(3):419–426.
16. Endsley JJ, et al. (2004) Characterization of bovine homologues of granulysin and NK-lysin. *J Immunol* 173(4):2607–2614.
17. Bickhart DM, et al. (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* 22(4):778–790.
18. Liu GE, et al. (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Res* 20(5):693–703.
19. Korbel JO, et al. (2008) The current excitement about copy-number variation: How it relates to gene duplications and protein families. *Curr Opin Struct Biol* 18(3):366–374.
20. Behe MJ, Snoke DW (2004) Simulating evolution by gene duplication of protein features that require multiple amino acid residues. *Protein Sci* 13(10):2651–2664.
21. Ohta T (1989) Role of gene duplication in evolution. *Genome* 31(1):304–310.

22. Elsik CG, et al.; Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324(5926):522–528.
23. Meade KG, Cormican P, Narciandi F, Lloyd A, O'Farrelly C (2014) Bovine β-defensin gene family: Opportunities to improve animal health? *Physiol Genomics* 46(1):17–28.
24. Scocchi M, Wang S, Zanetti M (1997) Structural organization of the bovine cathelicidin gene family and identification of a novel member. *FEBS Lett* 417(3):311–315.
25. Zanetti M (2004) Cathelicidins, multifunctional peptides of the innate immunity. *J Leukoc Biol* 75(1):39–48.
26. Larson JH, Marron BM, Beever JE, Roe BA, Lewin HA (2006) Genomic organization and evolution of the ULBP genes in cattle. *BMC Genomics* 7:227.
27. Walker AM, Roberts RM (2009) Characterization of the bovine type I IFN locus: Rearrangements, expansions, and novel subfamilies. *BMC Genomics* 10:187.
28. Irwin DM, Biegel JM, Stewart CB (2011) Evolution of the mammalian lysozyme gene family. *BMC Evol Biol* 11:166.
29. Neibergs HL, et al.; Bovine Respiratory Disease Complex Coordinated Agricultural Project Research Team (2014) Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned holstein calves. *BMC Genomics* 15(1):1164.
30. McVean G (2010) What drives recombination hotspots to repeat DNA in humans? *Philos Trans R Soc Lond B Biol Sci* 365(1544):1213–1218.
31. Stoppa-Lyonnet D, et al. (1991) Recombinational biases in the rearranged C1-inhibitor genes of hereditary angioedema patients. *Am J Hum Genet* 49(5):1055–1062.
32. Lehrman MA, et al. (1985) Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* 227(4683): 140–146.
33. Bailey JA, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73(4):823–834.
34. Bailey JA, et al. (2002) Recent segmental duplications in the human genome. *Science* 297(5583):1003–1007.
35. Cheng Z, et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437(7055):88–93.
36. Tuzun E, Bailey JA, Eichler EE (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res* 14(4):493–506.
37. She X, Cheng Z, Zöllner S, Church DM, Eichler EE (2008) Mouse segmental duplication and copy number variation. *Nat Genet* 40(7):909–914.
38. Nicholas TJ, et al. (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* 19(3):491–499.
39. Liu GE, et al. (2009) Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* 10:571.
40. Sharp AJ, et al. (2005) Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 77(1):78–88.
41. Graubert TA, et al. (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 3(1):e3.
42. Redon R, et al. (2006) Global variation in copy number in the human genome. *Nature* 444(7118):444–454.
43. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30(12):2725–2729.
44. Huddleston J, et al. (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* 24(4):688–696.
45. Parsons JD (1995) Miropeats: Graphical DNA sequence comparisons. *Comput Appl Biosci* 11(6):615–619.
46. Golosova O, et al. (2014) Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. *PeerJ* 2:e644.
47. Okonechnikov K, Golosova O, Fursov M; UGENE team (2012) Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 28(8):1166–1167.
48. Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.
49. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474.
50. Lawrence M, et al. (2013) Software for computing and annotating genomic ranges. *PLOS Comput Biol* 9(8):e1003118.
51. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80.
52. R Core Team (2015) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
53. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9(7):671–675.

Chapter 6

Retrotransposons: Genomic and Trans-Genomic Agents of Change

Throughout this thesis, accumulation of retrotransposons has been treated as if they are bound to the genome in which they reside. In this chapter horizontal transfer of retrotransposons is explored in depth. This chapter shows that retrotransposons are powerful agents of change and have the ability to alter genome evolution across species boundaries. The following excerpt appears as chapter 4 in *Evolutionary Biology: Biodiversification from Genotype to Phenotype* and discusses the role of retrotransposons as drivers of genome evolution.

Statement of Authorship

Title of Paper	Retrotransposons: Genomic and Trans-Genomic Agents of Change
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Adelson, D. L., Buckley, R. M., Ivancevic, A. M., Qu, Z., & Zeng, L. (2015). Retrotransposons: Genomic and Trans-Genomic Agents of Change. In *Evolutionary Biology: Biodiversification from Genotype to Phenotype* (pp. 55-75). Springer International Publishing.

Co-Author

Name of Co-Author (Candidate)	Reuben Buckley		
Contribution to the Paper	Designed and prepared several figures, provided suggestions and proof-read the manuscript.		
Overall percentage (%)	15%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the **second** author of this paper.		
Signature		Date	25/06/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Principal Author	David L. Adelson		
Contribution to the Paper	Wrote the manuscript		
Signature		Date	25/8/2017

Name of Co-Author	Atma M. Ivanncevic		
Contribution to the Paper	Designed and prepared figures, provided suggestions and proof-read the manuscript.		
Signature		Date	28/6/17

Name of Co-Author	Lu Zeng		
Contribution to the Paper	Designed and prepared figures, provided suggestions and proof-read the manuscript.		
Signature		Date	31/07/2017

Name of Co-Author	Zhipeng Qu		
Contribution to the Paper	Designed and prepared figures, provided suggestions and proof-read the manuscript.		
Signature		Date	31/07/2017

Chapter 4
Retrotransposons: Genomic and Trans-Genomic Agents of Change

David L. Adelson, Reuben M. Buckley, Atma M. Ivancevic, Zhipeng Qu and Lu Zeng

Abstract Genome structure in higher eukaryotes is highly dependent on the type and abundance of transposable elements, particularly retrotransposons, in their non-coding DNA. Retrotransposons are generally viewed as genomic parasites that must be suppressed in order to ensure genome integrity. This perception is based on the instances of retrotransposons having caused deleterious structural variation in genomes. Recent data are beginning to provide a more positive view of the impact of retrotransposons, particularly in mammals, where the evolution of the placenta has depended on the exaptation of a type of retrotransposon, endogenous retroviruses. Finally, exosome trafficking of retrotransposons between cells has been shown to induce the innate immune system gene expression, possibly indicative of a role for retrotransposons in the regulation of the innate immune system. It may be time for us to review the status of retrotransposons and reclassify them as symbionts rather than parasites.

4.1 Evolutionary Origin and Structure of Retrotransposons

Genome structure and function are two sides of the same coin, and retrotransposons (AKA retrotransposable elements, retroelements and retroposons), self-replicating DNA sequences that are found in all eukaryotic taxa, have the capacity to make larger changes to genome structure than other sources of variation—such as DNA polymerase errors that lead to single nucleotide variation (SNV). Because retrotransposons can account for the majority of the genome sequence in eukaryotes, their accumulation and clade specificity have been implicated in speciation, regulation of gene expression, exaptation and structural variation. Understanding the

D.L. Adelson (✉) · R.M. Buckley · A.M. Ivancevic · Z. Qu · L. Zeng
School of Biological Sciences, University of Adelaide, North Terrace, Adelaide, SA 5005, Australia
e-mail: david.adelson@adelaide.edu.au

P. Pontarotti (ed.), *Evolutionary Biology: Biodiversification from Genotype to Phenotype*, DOI 10.1007/978-3-319-19932-0_4

mechanisms that govern retrotransposon distribution and replication is thus of fundamental importance.

The evolutionary origin of retrotransposons is a matter of debate, but sequence similarity of their reverse transcriptases with the catalytic subunit of telomerase (Eickbush 1997; Lingner et al. 1997) and phylogenetic studies of reverse transcriptase sequences can be interpreted to indicate that reverse transcriptase may have evolved from telomerase, or telomerase is the result of co-opting reverse transcriptase. However, there are also good arguments for the ancient, prokaryotic origin of reverse transcriptase as a descendant of group II introns, which are mobile, self-splicing introns (Boeke 2003).

Retrotransposons can be divided into four major classes (Eickbush and Jamburuthugoda 2008). This classification is based on the reverse transcriptase enzyme required for replication and encoded by these elements. In vertebrates, retrotransposons can account for half of the genome sequence, and in plants, up to 70 % of the genome. This chapter is focused on the mammalian/vertebrate retrotransposons and these are commonly described as falling into two broad categories: those containing long terminal repeats (LTR) and those not containing LTR (non-LTR) (Jurka et al. 2007).

Non-LTR retrotransposons encode their own internal promoter and one or two open reading frames (ORFs) with reverse transcriptase and endonuclease activities that are used for replication (Fig. 4.1). LTR containing retrotransposons resemble (endogenous) retroviruses (ERVs) in that they can contain additional ORFs similar to those found in retroviruses, and these are referred to as endogenous retrovirus-like elements (ERVL). ERVL LTR retrotransposons are believed to have evolved from DNA transposons (Bao et al. 2010) and then acquired additional genes from viruses such as *env*, allowing them to become retrovirus-like and to produce infectious particles.

4.2 The Retrotransposon Life cycle

Retrotransposons replicate via an RNA intermediate that is reverse transcribed and reinserted into the genome (Fig. 4.1) at short target motifs (Fig. 4.2) (Cost and Boeke 1998). For non-LTR retrotransposons, also called long interspersed elements (LINE), transcription is initiated by an internal Pol II promoter and the resulting transcript is then translated to produce two proteins, one of which, ORF2p has both reverse transcriptase and endonuclease activities (Feng et al. 1996; Moran et al. 1996). ORF2p has the ability to recognise short target sequences and initiate nicks at those locations which subsequently serve to prime the reverse transcription of the retrotransposon RNA directly into the genome (Eickbush and Jamburuthugoda 2008; Morrish et al. 2002).

Some retrotransposons do not contain ORFs (non-autonomous) and are dependent on retrotransposons that do (autonomous) (Jurka et al. 2007). Autonomous retrotransposons are longer (LINEs), whereas the shorter, non-autonomous

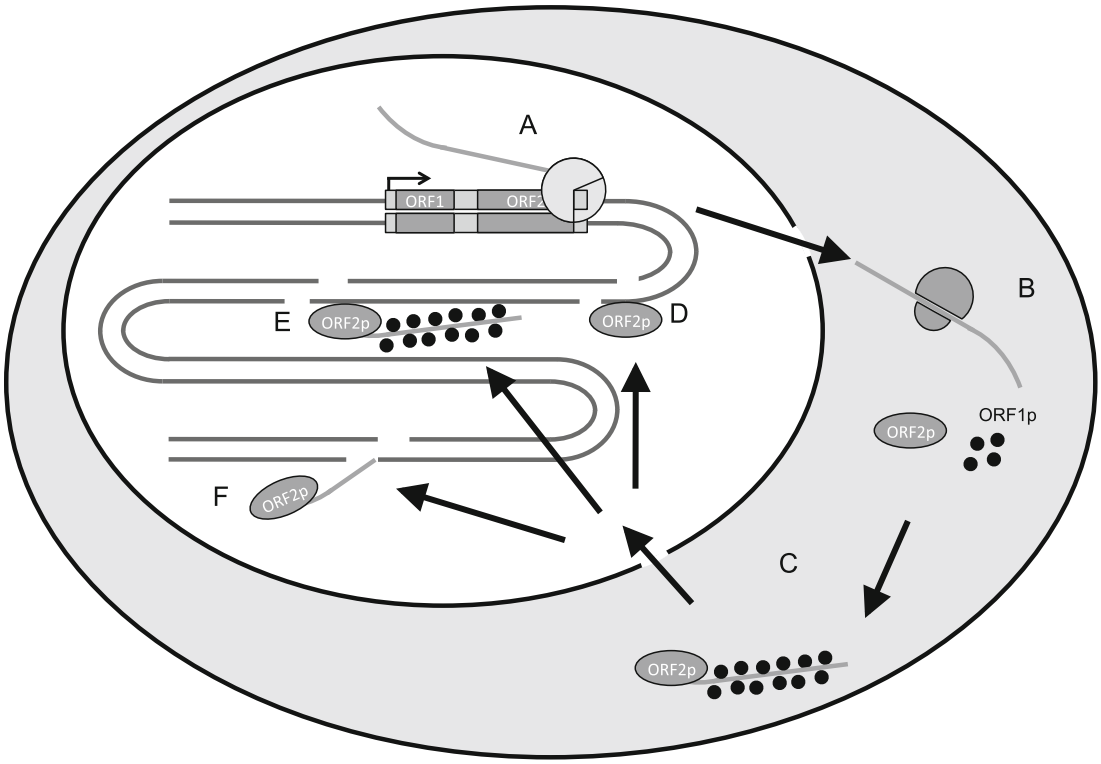


Fig. 4.1 Retrotransposon life cycle: *A* TEs are transcribed by RNA Pol II and exported to the cytoplasm (Swergold 1990). *B* In the cytoplasm, ORF1 and ORF2 are both translated. The ORF1 protein (*ORF1p*) is an RNA-binding protein believed to aid the entry of LINE L1 RNA into the nucleus (Martin 2006). The ORF2 protein (*ORF2p*) has both endonuclease and reverse transcriptase activities (Feng et al. 1996; Moran et al. 1996). *C* To enter the nucleus, ORF1p and ORF2p form a complex with the L1 RNA known as a ribonuclear protein (*RNP*) (Martin 2006). *D* The endonuclease activity of ORF2p creates double-stranded breaks without insertion of TEs (Gasior et al. 2006). *E* The endonuclease activity is essential for the process of target-primed reverse transcription (*TPRT*). TPRT requires that ORF2p creates a nick in each strand at the integration site. The LINE L1 RNA is then used as a template for the reverse transcriptase activity of ORF2p (Cost et al. 2002). *F* L1 RNA is able to insert into and aid in repairing double-stranded breaks independent of the endonuclease activity of ORF2p (Morrish et al. 2002)

elements are called short interspersed elements (SINEs). While LINEs are usually ubiquitously distributed across taxa, SINEs are usually clade specific, as they result from the fusion of an internal promoter containing transcript with the 3' end of a LINE.

The mechanism of SINE creation is still an open question, but most likely is a function of aspects of the LINE life cycle. SINEs have a composite structure: a 5' end similar to 5' tRNA, 7SL RNA or 5S rRNA promoters, a unique region and a 3' end similar to the 3' tail of LINEs (Piskurek and Jackson 2012). The most accepted hypothesis on SINE origins is based on the proposed template-switching mechanism of Buzdin et al. (Buzdin et al. 2002; Gilbert and Labuda 2000; Gogvadze and Buzdin 2009, Kramerov and Vassetzky 2005; Ohshima and Okada 2005). This template-switching mechanism is based on the study of pseudogenes, where the LINE (L1) reverse transcriptase switches from its own L1 mRNA to other nearby

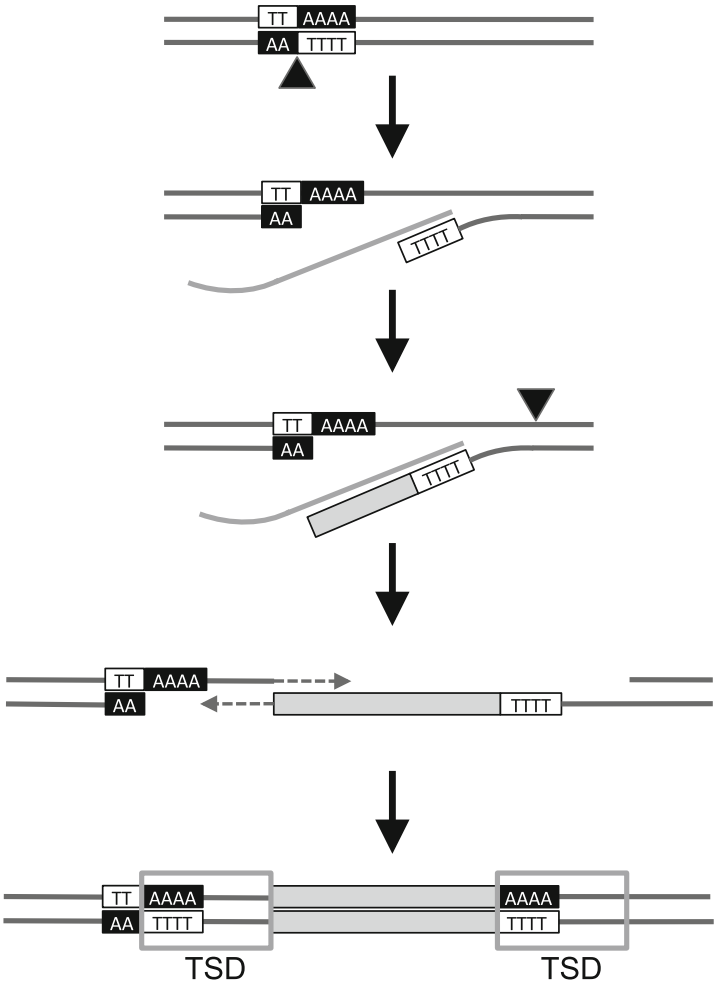


Fig. 4.2 Target-primed Reverse Transcription (*TPRT*) is how retrotransposons are inserted into the genome. ORF2p endonuclease activity creates a nick in the DNA at the AA/TTTT target site (Cost and Boeke, 1998). ORF2p reverse transcriptase activity then uses the cDNA copy as a template for DNA synthesis. Next ORF2p endonuclease activity creates a second nick in the DNA. The second DNA strand is then synthesised via double-strand break (*DSB*) repair and results in the formation of short target site duplications (*TSD*)

mRNA sequences through an RNA–RNA recombination process, thus creating new recombinant pseudogenes (and possibly SINEs) during L1 insertion (Buzdin et al. 2002; Gogvadze et al. 2007; Ichiyanagi et al. 2007; Piskurek and Jackson 2012). However, other investigators have suggested direct transposon into transposon (TnT) insertion as an alternative mechanism for the creation of novel transposable elements (Giordano et al. 2007; Ichiyanagi et al. 2007; Kriegs et al. 2007). The TnT mode of retrotransposon generation is what has led to the formation of SVA (SINE/VNTR/Alu) elements in humans, which are chimeric elements that can be mobilised by L1 elements and contain Alu-like sequence, Variable Number of Tandem Repeats (VNTR) sequence and SINE-R sequence resulting from a series of TnT events (Ostertag et al. 2003). The template-switching and TnT mechanisms are not mutually exclusive, and it is clear that both operate to create new SINEs, but at present we do not know which mechanism dominates.

Because retrotransposons can control their own expression through internal promoters [Pol II for LINEs and Pol III for SINEs and ERVs (Belancio et al. 2010a; Dieci et al. 2013)], expression is inextricably linked to the retrotransposon replication and to the evolution of new SINEs. As a result of this ability to autonomously insert new copies from expressed sequences into the genome, eukaryotes

have evolved mechanisms to keep retrotransposon expression in check in order to avoid large-scale deleterious structural variation.

4.2.1 Retrotransposon Suppression

There appear to be two main mechanisms for retrotransposon suppression: transcriptional repression and post-transcriptional degradation (Fig. 4.3). Transcriptional repression can be caused by methylation of retrotransposon promoters or alteration of chromatin state to make retrotransposons transcriptionally inaccessible. Proof for the importance of methylation is evident from the phenotype of *dnmt3l* (DNA (cytosine-5)-methyltransferase 3-like) knockout mice (Bourc'his and Bestor 2004; Webster et al. 2005), which undergo meiotic catastrophe associated with the rampant expression of retrotransposons in male germ cells. The *dnmt3l* locus encodes a protein that regulates methyl transferase activity required to methylate and suppress the activity of CpG islands in retrotransposon promoters (Vlachogiannis et al. 2015). In addition to CpG island methylation, transcription can be repressed by the alteration of chromatin status (Fadloun et al. 2013), and this may be mediated by piRNA transported to the nucleus (Kuramochi-Miyagawa et al. 2008).

Post-transcriptional degradation of retrotransposon RNA in the male germ line is mediated by piRNAs derived from retrotransposon sequences and amplified by the ping-pong reaction (Aravin et al. 2008). In the female germ line, the situation appears to be different, with siRNAs shown to mediate retrotransposon transcript destruction via the RNA-induced silencing complex (RISC) pathway (Ciaudo et al. 2013; Watanabe et al. 2008).

There may also be additional mechanisms that can suppress retrotransposons at the translational level (Grivna et al. 2006; Tanaka et al. 2011) or even at the post-translational level to interfere with ORF proteins binding to retrotransposon transcripts (Fig. 4.3) (Goodier et al. 2012). In spite of all of these mechanisms to suppress retrotransposons at various steps in their life cycle, they are still transcribed at some developmental stages and in many somatic tissues (Belancio et al. 2010b). Perhaps suppression is a loaded term in this context and perhaps what we are observing is actually the regulation of retrotransposon expression.

4.2.2 Retrotransposon Expression

At certain phases of the mammalian life cycle, retrotransposons are negatively regulated to a lesser degree and are therefore transcribed and able to retrotranspose. Because methylation of cytosine to 5-methyl-cytosine (5mC) is critical to retrotransposon silencing, retrotransposons are potentially most active at times of low genomic 5mC content, which occurs in mouse embryos at around 3.5 days of embryonic development and also in primordial germ cells (Hackett and Surani 2013).

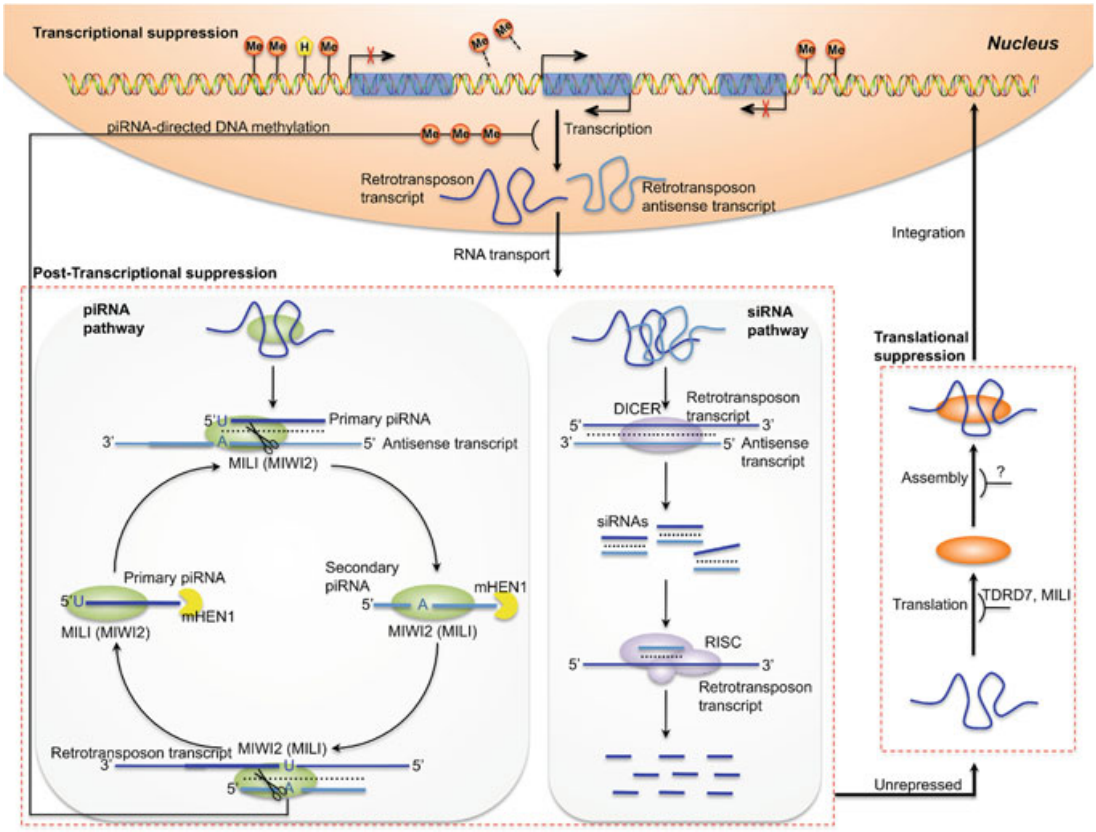


Fig. 4.3 A schematic overview of retrotransposon suppression. Retrotransposons can be suppressed by different mechanisms throughout their life cycle (Crichton et al. 2014). Transcriptional suppression: In most cell types, retrotransposons are in a repressed state due to high levels of DNA methylation or histone modifications (Fadloun et al. 2013; Meissner et al. 2008). In some specific developmental stages and cell types, some retrotransposon RNAs can be transcribed bidirectionally and transported from the nucleus to the cytoplasm (Fadloun et al. 2013). Post-transcriptional suppression: Retrotransposon RNAs can be silenced through the piRNA pathway (mostly in the male germ line) or siRNA pathway (mostly in the female germ line). The ping-pong cycle is a well-characterised model for piRNA synthesis. In the mouse, sense retrotransposon RNAs are processed into primary piRNAs. MILI (or MIWI2) is recruited to cleave antisense retrotransposon RNAs into secondary piRNAs with the guidance of primary piRNAs, and mHEN1 is used to subsequently methylate their 3' termini. Secondary piRNAs then bind with MIWI2 (or MILI) to cleave sense retrotransposon RNAs into primary piRNAs and close the loop of the ping-pong cycle (Aravin et al. 2008). piRNAs can also be transported to the nucleus to repress the transcription of retrotransposon by directing DNA methylation (Kuramochi-Miyagawa et al. 2008). For the siRNA pathway, sense and antisense retrotransposon transcripts can form double-strand RNAs, which are cleaved into double-strand siRNAs by DICER. Then, double-stranded siRNAs are unwound and loaded into the RISC to guide the degradation of retrotransposons (Ciaudo et al. 2013; Watanabe et al. 2008). Translational suppression: The Tudor domain-containing protein TDRD7 and MILI might be involved in the suppression of retrotransposon activity during translation (Grivna et al. 2006; Tanaka et al. 2011). Other repression mechanisms may also exist at later stages, such as the assembly stage of retrotransposon RNA and retrotransposon-encoded proteins (Goodier et al. 2012)

However, it is primarily in early embryos that L1 retrotransposons are transcribed and retrotranspose (Kano et al. 2009). Presumably, other suppression mechanisms keep retrotransposons in check in primary germ cells. In spite of significant levels of global 5mC in the genome at other stages of development, retrotransposons are also activated in specific somatic tissues, indicating that retrotransposon suppression is more complex than just ensuring high levels of 5mC, and it may be less stringent in some tissues/cell types. Faulkner et al. (2009) showed that up to 30 % of mouse or human transcripts from all tissues are of retrotransposon origin and that retrotransposons were transcribed in all tissues surveyed. Retrotransposon expression per se does not always mean that retrotransposition is occurring, as some retrotransposons have inserted into UTRs and are therefore transcribed as part of a mRNA. However, it has been shown in both neural progenitor cells and in the human brain that retrotransposition does occur at a detectable level, altering the genomic landscape of that tissue (Baillie et al. 2011; Coufal et al. 2009).

Retrotransposon expression and subsequent retrotransposition have significant impacts on the genomes of both germ line (via germ line insertions and early embryonic insertions) and soma. Germ line insertions can then be transmitted through vertical inheritance, while somatic insertions are not currently believed to contribute to the vertical inheritance of novel insertions. However, there is another mode of retrotransposon transmission: horizontal transfer, where retrotransposon sequences jump to another cell or species, and this type of transfer may be the result of a more general mechanism of intercellular retrotransposon transfer.

4.3 Horizontal Transfer

Horizontal transfer of transposons has been demonstrated in plants, insects and vertebrates. In the context of retroviruses (including ERVs that have maintained ORFs to support an infectious life cycle), horizontal transfer is a relatively commonplace event. For example, in plants, horizontal transfer of transposable elements is both widespread and frequent (El Baidouri et al. 2014). In animals, horizontal transfer of DNA transposons is also widespread (Ivancevic et al. 2013). A good example is in *Drosophila melanogaster* where P-elements swept through the population starting in the 1950s via horizontal transfer (Daniels et al. 1990). *Mariner* elements are also horizontally transmitted between species, including both insects and mammals (Lampe et al. 2003; Lohe et al. 1995; Maruyama and Hartl 1991). Furthermore, Space Invader (*SPIN*) elements have been horizontally transferred in mammals and other tetrapods, as have OC1 elements (Gilbert et al. 2010; Pace et al. 2008). It was not until the 1990s that the first evidence for horizontal transfer of retrotransposons was published, when the patchy phylogenetic distribution and likely horizontal transfer of BovB retrotransposons was first reported (Kordis and Gubensek 1998, 1999a).

4.3.1 BovB: An Example of Widespread Horizontal Transfer

The BovB retrotransposon (also known as LINE-RTE) is a 3.2 kb LINE with at least one large ORF encoding a reverse transcriptase and a possible small ORF1 overlapping with the large ORF (Malik and Eickbush 1998). In cattle and sheep, over a thousand full length BovB, hundreds of thousands of 5' truncated BovB fragments and derived SINEs (Bov-tA and Bov-tA2 (Lenstra et al. 1993; Okada and Hamada 1997) account for ~25 % of the genome sequence (Adelson et al. 2009; Jiang et al. 2014). The high degree of sequence conservation of BovB with sequences detected from the venom gland of *Vipera ammodytes* gave the first support to the idea of horizontal transfer of this retrotransposon (Kordis and Gubensek 1998, 1999b). BovB is now known to have a widespread, but patchy phylogenetic distribution, coupled to a high degree of sequence conservation, two of the hallmarks of horizontally transferred DNA (Fig. 4.4).

Even though BovB has horizontally transferred across a wide range of species, it has not always colonised the genome to the same extent in different species. Some

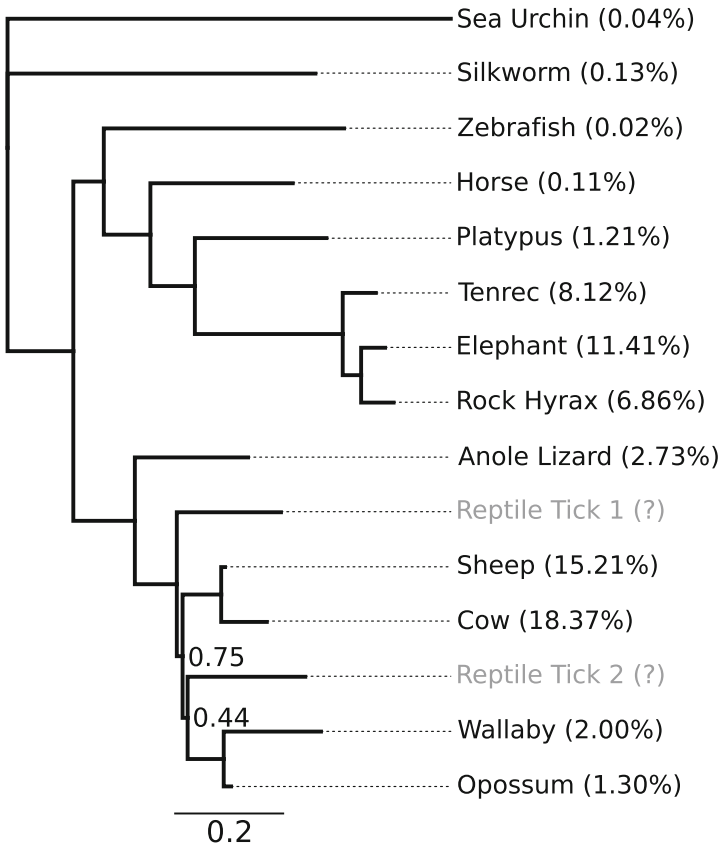


Fig. 4.4 BovB phylogeny Maximum likelihood tree of aligned BovB sequences based on Walsh et al. (2013), showing the sporadic distribution, sequence similarity and abundance of BovB elements across taxa. Local support values are only shown if <0.9. The labels at each branch tip give the species common name and (*in brackets*) the percentage of genome sequence identified as BovB elements for that species. Reptile Tick 1 is *Bothriocroton hydrosauri*, Reptile Tick 2 is *Amblyomma limbatum*; and the BovB genome coverage for these ticks is unknown

lineages such as ruminants and afrotheria have a high percentage of their genomes derived from BovB, whereas in other species BovB has not retrotransposed as prolifically (Fig. 4.4). This difference may be indicative of either variability in how different species suppress retrotransposons or it may simply reflect stochasticity in the population dynamics of retrotransposon expansion in different genomes. Presumably, the initial horizontal transfer event that results in retrotransposition and replication needs only a single germ line incorporation which can either replicate exponentially or "fizzle out" within the "genomic ecosystem" (Brookfield 2005; Le Rouzic et al. 2007). It is clear based on the currently available small and biased (towards mammals) sample of available genome sequences that retrotransposons as exemplified by BovB are capable of widespread and near ubiquitous horizontal transfer, and that this transfer might be enabled by parasites, such as ticks, that feed on blood. However, what is currently lacking is/are the molecular mechanism(s) for these transfers.

4.3.2 Possible Mechanisms/Modes of Transfer

A number of vectors, including arthropods, viruses, snails and DNA transposons, have been proposed for horizontal transfer, and the current state of knowledge was recently summarised by Ivancevic et al. (2013). It is relatively easy to see how a virus or transposon might act as a vector to package or transpose retrotransposons, but at the molecular level, it is not as obvious how eukaryotic vectors might effect the transfer of retrotransposon sequences between species, let alone into the germ line of another species.

4.3.2.1 Viruses as Vectors

For retrotransposons, the only example at present of a molecular virus vector is the taterapox virus (a dsDNA virus) which may have mediated transfer of Sauria SINE between reptiles and West African rodents (Piskurek and Okada 2007). This can be viewed as a highly unusual transfer, as a non-autonomous retrotransposon should not be as likely to colonise a new genome after transfer as an autonomous retrotransposon, such as a LINE. However, if cognate autonomous LINEs are present in both source and recipient species, a non-autonomous SINE could replicate effectively in the recipient species. RNA viruses have also been proposed as vectors of horizontal transfer for retrotransposons as they might package non-LTR retrotransposon transcripts inside infectious virus particles, but a tangible example for this type of transfer has yet to be demonstrated. Interestingly, *Mariner*-like DNA transposons are the plausible vectors for transfer of the CR1 retrotransposon in butterflies and moths (Sormacheva et al. 2012).

4.3.2.2 Endogenous Retroviruses/LTR Retrotransposons

As mentioned in Sect. 4.1, LTR retrotransposons are believed to have arisen from retrotransposons that acquired viral genes allowing them to become infectious, possibly leading to the evolution of retroviruses (Shimotohno and Temin 1981). In addition, waves of retroviral invasions into eukaryotic genomes have resulted in the formation of ERVs. While some ERVs have remained endogenous, occasionally they are able to become infectious and transfer to other genomes, where they can cause disease and eventually become domesticated. This is currently the case for a rodent ERV that has infected Koalas and is causing leukaemia in its new host while colonising the germ line as a new ERV (Tarlinton et al. 2006). Over time, domesticated retroviruses (ERVs) have contributed significantly to the genomic landscape of eukaryotes and have been co-opted into various aspects of eukaryotic biology (Feschotte and Gilbert 2012). In addition to this evolution of the capacity for horizontal transfer via infection, it is possible that retroviruses could package non-infectious non-LTR retrotransposons as a part of their viral payload. While there is no solid evidence for such transfer, exosomes/microvesicles are able to incorporate virus particles and transfer them to adjacent cells. This raises the question of whether exosomes can also transfer retrotransposon sequences directly.

4.3.2.3 Exosomes/Vesicles as Vectors

Exosomes are a class of membrane vesicle that has recently been shown to contain protein and RNA including miRNAs, piRNAs and retrotransposon sequences that they can transport from cell to cell (Batagov and Kurochkin 2013, Li et al. 2013; Skog et al. 2008; Valadi et al. 2007; Villarroya-Beltri et al. 2013; Yuan et al. 2009). Furthermore, exosome transport of Pol III-produced retrotransposon sequences has been specifically shown to regulate cancer therapy resistance pathways, including interferon-stimulated genes by direct activation of retinoid acid-inducible gene 1 (RIG-I) (Boelens et al. 2014). One of the hallmarks of Pol III transcripts is their 5' triphosphate group, which is recognised specifically by RIG-I as a trigger for activation. Pol III is responsible for the transcription of primarily housekeeping-type genes such as tRNAs and rRNAs, but it also transcribes many other loci, including SINEs that have originated from a fusion of Pol III promoter containing transcripts with LINE 3' sequences (Belancio et al. 2010b; Dieci et al. 2013). Because retrotransposons are known to be somatically expressed (see Sect. 4.2.2) in many tissues and cell types, they are likely to be present in exosomes exported by those cell types.

In the context of horizontal transfer, one can envision a number of potential scenarios for intercellular transport of retrotransposon sequences by exosomes (Fig. 4.5). Exosome-mediated transfer could allow transfer of retrotransposon sequences from a mammal or reptile to somatic cells of a parasite such as a tick through blood-borne exosomes. Within the tick, exosome-mediated transfer could then allow transmission to the germ line from the soma and eventual transmission back to other species used as food sources by that species of tick.

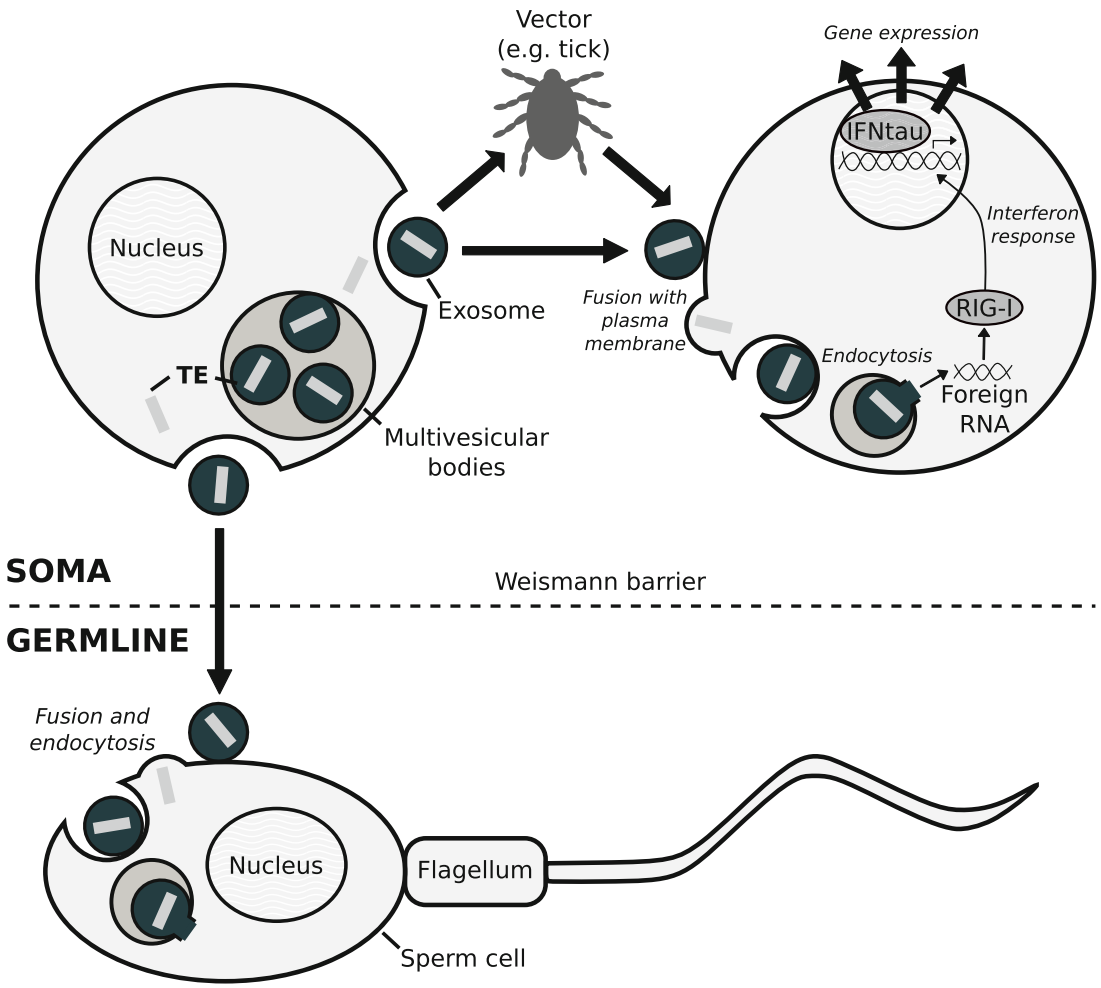


Fig. 4.5 Possible scenarios of intercellular transfer of transposable elements via exosomes. TEs packaged in exosomes can be transferred between both somatic and germline cells. Within an organism, a TE can travel from a somatic, exosome-generating cell directly (e.g. through the blood) into a somatic, exosome-target cell by fusing with the plasma membrane and undergoing endocytosis. Similarly, TEs can be horizontally transferred between the somatic cells of different organisms or species, via some kind of vector (e.g. a parasite). Exosomes can also carry TEs from the soma to the germ line, making them a permanent change in the genome that is eventually passed down to the offspring. Note that for simplicity only entry to the male germ line is shown above. In addition to the transfer of TEs, once inside the target cell, this "foreign RNA" from the TE can trigger an interferon pathway response by inducing the interferon signal transduction pathway via RIG-I. For example, in ruminants, exosomes loaded with ERV/TE RNAs trigger pattern recognition receptors, stimulating the innate immune system and production of interferon-tau, which plays a role in pregnancy recognition and placentation (see Sect. 4.4.4)

While one might envision that the existing piRNA-based suppression system might degrade these retrotransposon sequences rapidly, it also appears that retrotransposon sequences (as exosome cargo) have been co-opted into a signalling role for the innate immune system in vertebrates and used to activate interferon-stimulated genes in the absence of interferon (Dreux et al. 2012; Li et al. 2013). This would not be the first time that retrotransposon sequences have been co-opted for gene regulation (Feschotte 2008; Feschotte and Gilbert 2012), but it introduces a

new dimension of intercellular regulation of gene expression in the context of the evolutionary impact of retrotransposons.

4.4 Evolutionary Impacts

Retrotransposons are known to affect genome structure and hence function. The specific types of structural changes they introduce upon retrotransposition can have a wide-ranging set of subsequent effects in terms of genome structure, gene expression and gene function. More recently, it has become clear that retrotransposons have had a profound impact on the evolution of placentation in mammals.

4.4.1 Genome Structure

Retrotransposon insertion can directly perturb gene structure, but it can also have significant effects on a larger scale (Fig. 4.6). In particular, if retrotransposons form an array of elements with the same orientation on a chromosome, they can serve as

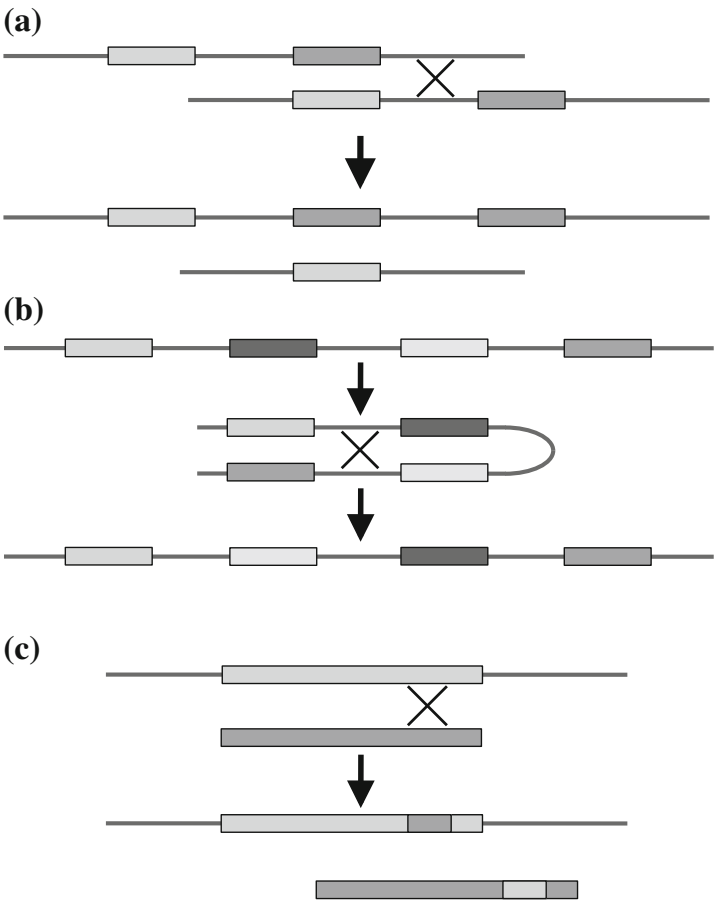


Fig. 4.6 Retrotransposons can lead to changes in genome structure. **a** Changes in CNVs result from non-allelic homologous recombination (*NAHR*) caused by the insertion of many TEs from the same family (Stankiewicz and Lupski 2002; Startek et al. 2015). **b** Chromosomal inversion is also the result of NAHR (Stankiewicz and Lupski 2002). **c** SINE elements have potential to drive change through gene conversion (Roy et al. 2000)

a substrate for non-allelic homologous recombination (NAHR) leading to segmental duplication (Fig. 4.6a) (Stankiewicz and Lupski 2002; Startek et al. 2015). However, statistical analysis of repeats in flanking regions of segmental duplications found that only ~10 % of segmental duplications could be attributed to flanking repetitive elements (Zhou and Mishra 2005). Other types of rearrangements have been shown to result from arrays of repeats such as inversions (Fig. 4.6b) and gene conversion (Fig. 4.6c).

While it is clear that retrotransposons can have indirect effects on genome structure as mentioned above, given the limitations inherent in identifying small segmental duplications and copy number variants the precise magnitude of these effects is unknown.

4.4.2 Gene Expression

As shown in Fig. 4.7, transposable elements can insert into and next to genes, affecting gene expression through multiple mechanisms, including epigenetic silencing of transcription, shortening a transcript via premature poly-Adenylation,

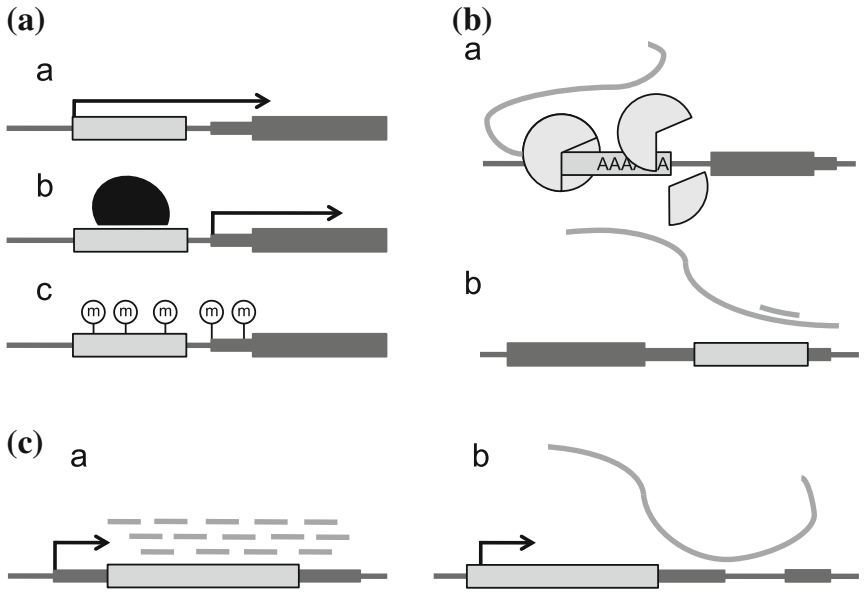


Fig. 4.7 Retrotransposons can alter gene expression. **a** 5' insertion of a retrotransposon with respect to a gene. *a* TEs are able to act as alternative promoters to adjacent genes (Faulkner et al. 2009; Speek 2001). *b* TEs are able to act as transcription factor binding sites (*TFBS*) and are thereby able to modulate gene expression (Bourque et al. 2008). *c* In plants, epigenetic silencing of TEs silences nearby genes; this is also likely to occur in animals (Buckley and Adelson 2014; Hollister and Gaut 2009). **b** 3' insertion of a retrotransposon *a* polyA signal/tail of the retrotransposon can result in shortened transcripts (Lee et al. 2008; Perepelitsa-Belancio and Deininger 2003). *b* Retrotransposon insertion in the 3' UTR of a gene can provide a target site for piRNAs which down-regulate gene expression (Watanabe et al. 2014). **c** Intergenic insertion of TEs. *a* Insertion of TEs into a piRNA cluster results in piRNAs that can target genes carrying TE-derived sequences (Yamamoto et al. 2013). *b* TEs involved in the origin and evolution of lncRNA (Kapusta et al. 2013)

driving piRNA expression or altering 3' UTR structure to affect mRNA stability. Analysis of retrotransposon insertions into or near genes has shown that many genes have been altered in ways that are likely to alter expression (Jjingo et al. 2011; Jordan et al. 2003) and analysis of enhancers has shown that retrotransposons drive the evolution of eukaryotic enhancers (McDonald et al. 1997). All of these effects on gene expression are subject to selection and are therefore part of the evolutionary process. Not all insertions into genes will affect regulation of gene expression, some can directly affect the coding sequence or coding potential of genes through exaptation.

4.4.3 Exaptation

When retrotransposons contribute to non-coding or protein coding exon sequences, they are referred to as exaptations. These exaptations may or may not be subject to immediate purifying selection, depending on the type of change they cause. Some exaptations that prove beneficial are selected for, but these are rare. Many examples of exaptation come from non-coding transcripts, where retrotransposon insertions have led to novel piRNA and miRNA transcripts (Jurka et al. 2007; Yamamoto et al. 2013). In fact, only ~50 instances of coding sequences derived from LTR retrotransposons syntenic between human and mouse have been identified (Jurka et al. 2007). One of these encodes the PEG10 (paternally expressed gene 10) locus, which is required for placentation. Occasionally, insertion of a retrotransposon sequence into an intron can lead to exonisation of part of the retrotransposon sequence as an alternative transcript through the presence of splice donor/acceptor sites in the sequence (Fig. 4.8). When this happens, sometimes the alternative transcripts are deleterious because of impaired function, and the regulation of alternative splicing may then become an additional regulatory mechanism for the affected gene (Lorenz et al. 2007).

4.4.4 Innate Immunity/Pregnancy Recognition

Some exaptations of retrotransposon sequences have been well-characterised, particularly in terms of the evolution of placentation. There is strong evidence for exaptation of ERV genes in both mouse and hominoid primates required for placental function (Chuong 2013; Haig 2012; Mallet et al. 2004). One of the most striking such exaptations is the role of endogenous jaagsiekte retrovirus (enJSRV) in ruminant pregnancy recognition and placentation. The domestic ruminant conceptus expresses interferon-tau (IFNT) from days 10 to 12, which dramatically alters gene expression in the uterine epithelium and stroma (Bazer et al. 2008; Dunlap et al. 2006; Gray et al. 2006; Spencer and Bazer 1995). At the same time, enJSRVs are released into the ruminant reproductive tract and they are known to

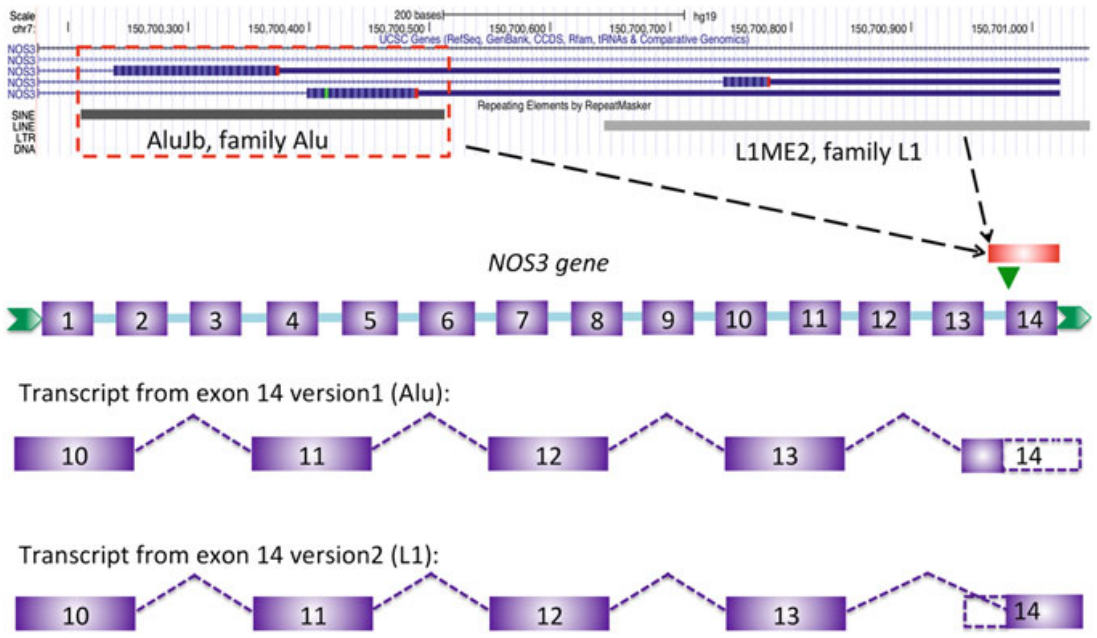


Fig. 4.8 Retrotransposon exaptation influences mRNA processing and can cause multiple splice variants. At the top, the UCSC browser (Kent et al. 2002) track for the human NOS3 gene is shown, including repeat element annotation. *Below*, a schematic of the 3' end of the human NOS3 gene illustrating an Alu element (*black bar*) inserted into intron 13. This retrotransposon provides exon 14 alternative splicing version 1. An adjacent L1 insertion can result in exon 14 alternative splicing version 2 (Lorenz et al. 2007). Dashed lines indicate a splicing event

regulate key peri-implantation development in the embryo and placenta (Dunlap et al. 2005, 2006). enJSRVs therefore have been exapted to regulate key aspects of development associated with implantation and placentation by virtue of their ability to trigger expression of IFNT expression in the conceptus. Recently, exosomes have been shown to be part of the specific mechanism used to trigger IFNT expression in this system, but without specifically testing for retrotransposon RNA content (Ruiz-Gonz ez et al. 2014, 2015). We speculate that exosomes loaded with retrotransposon sequences may also be involved in pregnancy recognition more generally in order to activate the STAT1 pathway in an interferon-free fashion.

SINE/ERV transcripts packaged into exosomes can trigger RIG-I in target cells leading to IFN independent activation of the IFN pathway, leading us to speculate that the role of retrotransposons is broader than previously thought, and that they may be involved in global regulation of the innate immune system.

4.5 Conclusion

Retrotransposons are abundant, found in a broad phylogenetic distribution and yet in spite of clade specific non-autonomous variants, exhibit a significant degree of commonality. Furthermore, their transcription is highly regulated, rather than

suppressed at all times. These facts, along with the evidence of pervasive and widespread horizontal transfer and an exosome-based mechanism for transfer that has likely co-evolved with the innate immune system and placentation, suggest to us that retrotransposons are not genomic parasites but rather genomic symbionts. We hypothesise that mammals and other vertebrates depend on these symbionts for cell-to-cell signalling in innate immunity and reproduction.

Acknowledgments The authors wish to thank R. Daniel Kortschak and Joy M. Raison for helpful discussions and advice.

References

Adelson DL, Raison JM, Edgar RC (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci USA 106:12855

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 31:785

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nat Cell Biol 479:534

Bao W, Kapitonov VV, Jurka J (2010) Ginger dna transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA 1(1):3. doi:10.1186/1759-8753-1-3

Batagov AO, Kurochkin IV (2013) Exosomes secreted by human cells transport largely mrna fragments that are enriched in the 3'-untranslated regions. Biol Direct 8:12. doi:10.1186/1745-6150-8-12

Bazer FW, Burghardt RC, Johnson GA, Spencer TE, Wu G (2008) Interferons and progesterone for establishment and maintenance of pregnancy: interactions among novel cell signaling pathways. Reprod Biol 8(3):179–211

Belancio VP, Roy-Engel AM, Deininger PL (2010a) All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 20(4):200–210. doi:10.1016/j.semcancer.2010.06.001

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P (2010 b) Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res 38:3909

Boeke JD (2003) The unusual phylogenetic distribution of retrotransposons: a hypothesis. Genome Res 13(9):1975–1983. doi:10.1101/gr.1392003

Boelens MC, Wu TJ, Nabet BY, Xu B, Qiu Y, Yoon T, Azzam DJ, Twyman-Saint Victor C, Wiemann BZ, Ishwaran H, Ter Brugge PJ, Jonkers J, Slingerland J, Minn AJ (2014) Exosome transfer from stromal to breast cancer cells regulates therapy resistance pathways. Cell 159 (3):499–513. doi:10.1016/j.cell.2014.09.051

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3 l. Nature 431(7004):96–99. doi:10.1038/nature02886

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18(11):1752–1762. doi:10.1101/gr.080663.108

Brookfield JFY (2005) The ecology of the genome—mobile dna elements and their hosts. Nat Rev Genet 6(2):128–136. doi:10.1038/nrg1524

Buckley RM, Adelson DL (2014) Mammalian genome evolution as a result of epigenetic regulation of transposable elements. Biomol Concepts 5(3):183–194. doi:10.1515/bmc-2014-0013

Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, Sverdlov E (2002) A new family of chimeric retrotranscripts formed by a full copy of u6 small nuclear rna fused to the 3' terminus of l1. Genomics 80(4):402–406

Chuong EB (2013) Retroviruses facilitate the rapid evolution of the mammalian placenta. Bioessays 35:853

Ciaudo C, Jay F, Okamoto I, Chen CJ, Sarazin A, Servant N, Barillot E, Heard E, Voinnet O (2013) Rnai-dependent and independent control of line1 accumulation and mobility in mouse embryonic stem cells. PLoS Genet 9(11):e1003,791. doi:10.1371/journal.pgen.1003791

Cost GJ, Boeke JD (1998) Targeting of human retrotransposon integration is directed by the specificity of the l1 endonuclease for regions of unusual dna structure. Biochemistry 37 (51):18081–18093

Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human l1 element target-primed reverse transcription in vitro. EMBO J 21(21):5899–5910

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH (2009) L1 retrotransposition in human neural progenitor cells. Nat Cell Biol 460:1127

Crichton JH, Dunican DS, Maclennan M, Meehan RR, Adams IR (2014) Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. Cell Mol Life Sci 71(9):1581–1605. doi:10.1007/s00018-013-1468-0

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A (1990) Evidence for horizontal transmission of the p-transposable element between drosophila species. Genetics 124:339

Dieci G, Conti A, Pagano A, Carnevali D (2013) Identification of rna polymerase iii-transcribed genes in eukaryotic genomes. Biochim Biophys Acta 1829(3–4):296–305. doi:10.1016/j.bbagrm.2012.09.010

Dreux M, Garaigorta U, Boyd B, Décembre E, Chung J, Whitten-Bauer C, Wieland S, Chisari FV (2012) Short-range exosomal transfer of viral rna from infected cells to plasmacytoid dendritic cells triggers innate immunity. Cell Host Microbe 12(4):558–570. doi:10.1016/j.chom.2012.08.010

Dunlap KA, Palmarini M, Adelson DL, Spencer TE (2005) Sheep endogenous betaretroviruses (enjsrvs) and the hyaluronidase 2 (hyal2) receptor in the ovine uterus and conceptus. Biol Reprod 73(2):271–279. doi:10.1095/biolreprod.105.039776

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, Spencer TE (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103(39):14390–14395. doi:10.1073/pnas.0603836103

Eickbush TH (1997) Telomerase and retrotransposons: which came first? Science (New York, NY) 277(5328):911–912

Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221

El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24(5):831–838. doi:10.1101/gr.164400.113

Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla ME (2013) Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of line-1 by rna. Nat Struct Mol Biol 20(3):332–338. doi:10.1038/nsmb.2495

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41:563

Feng Q, Moran J, Kazazian H, Boeke J (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905

Feschotte C (2008) Opinion—transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397
Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13(4):283–296. doi:10.1038/nrg3199
Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. Journal of Molecular Biology
Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1352
Gilbert N, Labuda D (2000) Evolutionary inventions and continuity of core-sines in mammals. J Mol Biol 298(3):365–377. doi:10.1006/jmbi.2000.3695
Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, Warburton P (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol 3: e137
Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66(23):3727–3742. doi:10.1007/s00018-009-0107-2
Gogvadze E, Barbisan C, Lebrun MH, Buzdin A (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus magnaporthe grisea suggests double template jumps during long interspersed nuclear element (line) reverse transcription. BMC Genomics 8:360. doi:10.1186/1471-2164-8-360
Goodier JL, Cheung LE, Kazazian HH Jr (2012) Mov10 rna helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 8(10):e1002941. doi:10.1371/journal.pgen.1002941
Gray CA, Abbey CA, Beremand PD, Choi Y, Farmer JL, Adelson DL, Thomas TL, Bazer FW, Spencer TE (2006) Identification of endometrial genes regulated by early pregnancy, progesterone, and interferon tau in the ovine uterus. Biol Reprod 74(2):383–394. doi:10.1095/biolreprod.105.046656
Grivna ST, Pyhtila B, Lin H (2006) Miwi associates with translational machinery and piwi-interacting rnas (pirnas) in regulating spermatogenesis. Proc Natl Acad Sci USA 103 (36):13415–13420. doi:10.1073/pnas.0605506103
Hackett JA, Surani MA (2013) Dna methylation dynamics during the mammalian life cycle. Philos Trans R Soc Lond B Biol Sci 368(1609):20110328. doi:10.1098/rstb.2011.0328
Haig D (2012) Retroviruses and the placenta. Current biology: CB
Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19 (8):1419–1428. doi:10.1101/gr.091678.109
Ichiyanagi K, Nakajima R, Kajikawa M, Okada N (2007) Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. Genome Res 17:33
Ivancevic AM, Walsh AM, Kortschak RD, Adelson DL (2013) Jumping the fine LINE between species: horizontal transfer of transposable elements in animals catalyses genome evolution. Bioessays 35:12
Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W, Stanton JA, Brauning R, Barris WC, Hourlier T, Aken BL, Searle SMJ, Adelson DL, Bian C, Cam GR, Chen Y, Cheng S, DeSilva U, Dixen K, Dong Y, Fan G, Franklin IR, Fu S, Fuentes-Utrilla P, Guan R, Highland MA, Holder ME, Huang G, Ingham AB, Jhangiani SN, Kalra D, Kovar CL, Lee SL, Liu W, Liu X, Lu C, Lv T, Mathew T, McWilliam S, Menzies M, Pan S, Robelin D, Servin B, Townley D, Wang W, Wei B, White SN, Yang X, Ye C, Yue Y, Zeng P, Zhou Q, Hansen JB, Kristiansen K, Gibbs RA, Flicek P, Warkup CC, Jones HE, Oddy VH, Nicholas FW, McEwan JC, Kijas JW, Wang J, Worley KC, Archibald AL, Cockett N, Xu X, Wang W, Dalrymple BP (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344(6188):1168–1173. doi:10.1126/science.1252806
Jjingo D, Huda A, Gundapuneni M, Mariño-Ramrez L, Jordan IK (2011) Effect of the transposable element environment of human genes on gene length and expression. Genome Biol Evol 3:259–271. doi:10.1093/gbe/evr015
Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Ann Rev Genomics Hum Genet

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH (2009) L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev 23:1303

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at ucsc. Genome Res 12(6):996–1006. doi:10.1101/gr.229102. Article published online before print in May

Kordis D, Gubensek F (1998) Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proc Natl Acad Sci USA 95(18):10704–10709

Kordis D, Gubensek F (1999a) Horizontal transfer of non-LTR retrotransposons in vertebrates. Genetica 107:121

Kordis D, Gubensek F (1999b) Molecular evolution of bov-b lines in vertebrates. Gene 238 (1):171–178

Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247:165–221. doi:10.1016/S0074-7696(05)47004-7

Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (aves: Galliformes). BMC Evol Biol 7:190. doi:10.1186/1471-2148-7-190

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, Hata K, Li E, Matsuda Y, Kimura T, Okabe M, Sakaki Y, Sasaki H, Nakano T (2008) Dna methylation of retrotransposon genes is regulated by piwi family members mili and miwi2 in murine fetal testes. Genes Dev 22(7):908–917. doi:10.1101/gad.1640708

Lampe DJ, Witherspoon DJ, Soto-Adames FN, Robertson HM (2003) Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol Biol Evol 20(4):554–562. doi:10.1093/molbev/msg069

Le Rouzic A, Boutin TS, Capy P (2007) Long-term evolution of transposable elements. Proc Natl Acad Sci USA 104(49):19375–19380. doi:10.1073/pnas.0705238104

Lee JY, Ji Z, Tian B (2008) Phylogenetic analysis of mrna polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic Acids Res 36(17):5581–5590. doi:10.1093/nar/gkn540

Lenstra JA, van Boxtel JA, Zwaagstra KA, Schwerin M (1993) Short interspersed nuclear element (sine) sequences of the bovidae. Anim Genet 24(1):33–39

Li CCY, Eaton SA, Young PE, Lee M, Shuttleworth R, Humphreys DT, Grau GE, Combes V, Bebawy M, Gong J, Brammah S, Buckland ME, Suter CM (2013) Glioma microvesicles carry selectively packaged coding and non-coding rnas which alter gene expression in recipient cells. RNA Biol 10(8):1333–1344. doi:10.4161/rna.25281

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. Science (New York, NY) 276:561

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol Biol Evol 12 (1):62–72

Lorenz M, Hewing B, Hui J, Zepp A, Baumann G, Bindereif A, Stangl V, Stangl K (2007) Alternative splicing in intron 13 of the human enos gene: a potential mechanism for regulating enos activity. FASEB J 21(7):1556–1564. doi:10.1096/fj.06-7434com

Malik H, Eickbush T (1998) The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol 15:1123

Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci USA 101:1731

Martin SL (2006) The orf1 protein encoded by line-1: structure and function during l1 retrotransposition. J Biomed Biotechnol 2006(1):45621. doi:10.1155/JBB/2006/45621

Maruyama K, Hartl DL (1991) Evidence for interspecific transfer of the transposable element mariner between Drosophila and Zaprionus. J Mol Evol 33:514

McDonald JF, Matyunina LV, Wilson S, Jordan IK, Bowen NJ, Miller WJ (1997) Ltr retrotransposons and the evolution of eukaryotic enhancers. Genetica 100(1–3):3–13

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale dna methylation maps of pluripotent and differentiated cells. Nature 454(7205):766–770. doi:10.1038/nature07107

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87(5):917–927

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV (2002) Dna repair mediated by endonuclease-independent line-1 retrotransposition. Nat Genet 31(2):159–165. doi:10.1038/ng898

Ohshima K, Okada N (2005) Sines and lines: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res 110(1–4):475–490. doi:10.1159/000084981

Okada N, Hamada M (1997) The 3' ends of trna-derived sines originated from the 3' ends of lines: a new example from the bovine genome. J Mol Evol 44(1):52–56

Ostertag E, Goodier J, Zhang Y, Kazazian H (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73:1444

Pace JK, Gilbert C, Clark MS, Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci USA 105:17023

Perepelitsa-Belancio V, Deininger P (2003) Rna truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 35(4):363–366. doi:10.1038/ng1269

Piskurek O, Jackson DJ (2012) Transposable elements: from dna parasites to architects of metazoan evolution. Genes (Basel) 3(3):409–422. doi:10.3390/genes3030409

Piskurek O, Okada N (2007) Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc Natl Acad Sci USA 104(29):12046–12051. doi:10.1073/pnas.0700531104

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL (2000) Potential gene conversion and source genes for recently integrated alu elements. Genome Res 10(10):1485–1495

Ruiz-González I, Xu J, Wang X, Burghardt RC, Dunlap K, Bazer FW (2014) Exosomes, endogenous retroviruses and toll-like receptors: pregnancy recognition in ewes. Reproduction

Ruiz-González I, Minten M, Wang X, Dunlap K, Bazer FW (2015) Involvement of TLR7 and TLR8 in conceptus development and establishment of pregnancy in Sheep. Reproduction

Shimotohno K, Temin HM (1981) Evolution of retroviruses from cellular movable genetic elements. Cold Spring Harb Symp Quant Biol 45(Pt 2):719–730

Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Sena-Esteves M, Curry WT Jr, Carter BS, Krichevsky AM, Breakefield XO (2008) Glioblastoma microvesicles transport rna and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol 10(12):1470–1476. doi:10.1038/ncb1800

Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O (2012) Vertical evolution and horizontal transfer of cr1 non-ltr retrotransposons and tc1/mariner dna transposons in lepidoptera species. Mol Biol Evol 29(12):3685–3702. doi:10.1093/molbev/mss181

Speek M (2001) Antisense promoter of human l1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21(6):1973–1985. doi:10.1128/MCB.21.6.1973-1985.2001

Spencer TE, Bazer FW (1995) Temporal and spatial alterations in uterine estrogen receptor and progesterone receptor gene expression during the estrous cycle and early pregnancy in the ewe. Biol Reprod 53(6):1527–1543

Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18(2):74–82

Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A (2015) Genome-wide analyses of line-line-mediated nonallelic homologous recombination. Nucleic Acids Res. doi:10.1093/nar/gku1394

Swergold GD (1990) Identification, characterization, and cell specificity of a human line-1 promoter. Mol Cell Biol 10(12):6718–6729

Tanaka T, Hosokawa M, Vagin VV, Reuter M, Hayashi E, Mochizuki AL, Kitamura K, Yamanaka H, Kondoh G, Okawa K, Kuramochi-Miyagawa S, Nakano T, Sachidanandam R, Hannon GJ, Pillai RS, Nakatsuji N, Chuma S (2011) Tudor domain containing 7 (tdrd7) is essential for dynamic ribonucleoprotein (rnp) remodeling of chromatoid bodies during spermatogenesis. Proc Natl Acad Sci USA 108(26):10579–10584. doi:10.1073/pnas.1015447108

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442 (7098):79–81. doi:10.1038/nature04841

Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO (2007) Exosome-mediated transfer of mrnas and micrornas is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9(6):654–659. doi:10.1038/ncb1596

Villarroya-Beltri C, Gutiérrez-Vázquez C, Sánchez-Cabo F, Pérez-Hernández D, Vázquez J, Martin-Cofreces N, Martinez-Herrera DJ, Pascual-Montano A, Mittelbrunn M, Sánchez-Madrid F (2013) Sumoylated hnrnpa2b1 controls the sorting of mirnas into exosomes through binding to specific motifs. Nat Commun 4:2980. doi:10.1038/ncomms3980

Vlachogiannis G, Niederhuth CE, Tuna S, Stathopoulou A, Viiri K, de Rooij DG, Jenner RG, Schmitz RJ, Ooi SKT (2015) The dnmt3 l add domain controls cytosine methylation establishment during spermatogenesis. Cell Rep. doi:10.1016/j.celrep.2015.01.021

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA 110:1012

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H (2008) Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. Nature 453(7194):539–543. doi:10.1038/nature06908

Watanabe T, Cheng EC, Zhong M, Lin H (2014) Retrotransposons and pseudogenes regulate mrnas and lncrnas via the pirna pathway in the germline. Genome Res. doi:10.1101/gr.180802.114

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ, Antonarakis SE, de Kretser DM, Hedger MP, Peterson P, Carroll BJ, Scott HS (2005) Meiotic and epigenetic defects in dnmt3l-knockout mouse spermatogenesis. Proc Natl Acad Sci USA 102(11):4068–4073. doi:10.1073/pnas.0500702102

Yamamoto Y, Watanabe T, Hoki Y, Shirane K, Li Y, Ichiiyanagi K, Kuramochi-Miyagawa S, Toyoda A, Fujiyama A, Oginuma M, Suzuki H, Sado T, Nakano T, Sasaki H (2013) Targeted gene silencing in mouse germ cells by insertion of a homologous dna into a pirna generating locus. Genome Res 23(2):292–299. doi:10.1101/gr.137224.112

Yuan A, Farber EL, Rapoport AL, Tejada D, Deniskin R, Akhmedov NB, Farber DB (2009) Transfer of micrornas by embryonic stem cell microvesicles. PLoS ONE 4(3):e4722. doi:10.1371/journal.pone.0004722

Zhou Y, Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci USA 102(11):4051–4056. doi:10.1073/pnas.0407957102

Chapter 7

Conclusions and Future Directions

Retrotransposons are a powerful force in mammalian genome evolution, making up a large fraction of mammalian non-coding DNA. However, their large-scale accumulation patterns and interactions with genome structure upon insertion are poorly characterised

Throughout my thesis I analysed the factors that shape retrotransposon accumulation as well as their potential impact on genome evolution. I developed novel comparative genomics approaches that made it possible to capture complex evolutionary events across large sections of poorly conserved non-coding DNA. My results showed that L1s and SINEs accumulated along similar evolutionary trajectories consistent with genome-wide chromatin structure and gene density. Retrotransposons insert into open chromatin, which is usually found in gene-rich genomic regions. In these regions smaller insertions are much more likely to be tolerated. Therefore, after insertion, L1s are usually purged from gene-rich regions due to purifying selection, while tolerated insertions in gene-poor regions accumulate over time. These retrotransposon accumulation dynamics are conserved across mammals and result in significant overlap of L1 and SINE enriched regions across different species.

While L1s and SINEs tend to accumulate in the same regions independently, it is important to realise that conserved dynamics across different systems can occasionally lead to divergent outcomes. From analysing the regional DNA gain and loss in human and mouse, I found that regional rates of DNA turnover were consistent with divergent evolution of genome architecture. Initially this seemed contrary to the above findings, however closer analysis showed that this was mostly caused by lineage-specific accumulation rates of L1s and SINEs. Moreover, DNA loss occurred in open chromatin where retrotransposons tend to insert, indicating that particular genomic regions undergo a high degree of ‘churning’.

From the approaches I developed in this thesis, there are two predominant methods:

humanisation of retrotransposon distributions and identification of DNA gain and loss events. Humanisation of retrotransposon distributions is the mapping of the retrotransposon content of large genomic regions (approximately 1 Mb) from a non-human species over to the human genome, essentially modelling non-human retrotransposon accumulation as if it had occurred within human chromosomes. The novelty of this approach meant that for the first time I was able to directly compare retrotransposon accumulation patterns on a region to region basis across many species simultaneously. This was a significant improvement on previous analyses that mainly relied on some other genomic feature such as GC content as a proxy for comparing retrotransposon accumulation patterns. For identification of gain and loss events, I took advantage of genome-wide pairwise alignments from multiple species. Similar to previous approaches I used retrotransposons and outgroup ancestry to assign gaps between a reference and query genome as either DNA gains or DNA losses. However, my method extended this approach by utilising the multi-level annotation of genome alignment nets, making it possible to untangle complex evolutionary genomic rearrangements and map individual events between species. Perhaps the most novel aspect of this method is the ability to both map and quantify DNA loss events at the loci where they originally occurred. Previously, DNA loss events from a particular genome had only been analysed within the genomes of other species, where the DNA still remained. The methods developed in this thesis were extremely effective at illuminating the hidden complexity of mammalian genome evolution and the impact of retrotransposons.

One of the primary limitations of this thesis is the breadth of analysed species. Until recently, mammalian genome sequencing was an expensive endeavour, causing many of the earlier projects to be centred around species of medical or agricultural importance. This is problematic for studying evolution, as many of these species have experienced artificial evolutionary scenarios such as domestication. These kinds of processes can result in species carrying unique genomic signatures that may be uncharacteristic of evolution in the wild. As sequencing projects with new technologies lead to a broad range of high quality genomes, it will be possible to characterise a much wider selection of mammalian genomes.

Another limitation for my analysis was identification and classification of retrotransposons. Due to their size and repetitive nature, retrotransposons are difficult to map as full length elements, as they are often longer than the reads generated from most sequencing platforms. For many genomes high retrotransposon density can cause genome misassembly. Additionally, many programs used to identify retrotransposons use a library based approach, where retrotransposons are identified based on alignment with known sequences.

Because most work focuses on human and mouse, their retrotransposons are classified into very specific families. In contrast, other species retrotransposon classifications are much less specific. This makes it difficult to associate individual retrotransposon families with species divergence.

In this thesis there are many areas where further analysis can be performed, expanding the scope of my findings. Throughout my research I characterised novel approaches for analysing retrotransposon mediated genome evolution, such as using pairwise alignments to detect DNA gain and loss events. Applying DNA gain and loss detection to a wider range of species will help develop further insight into the evolutionary impact of retrotransposons. One situation where DNA gain and loss detection would be useful is in species that have a divergent retrotransposon landscape from the species I analysed in chapters 2 and 3. For example, there are various mammalian clades where BovB elements have been introduced through independent horizontal transfer events. In these clades BovBs and their associated SINEs add an extra dimension to retrotransposon accumulation dynamics; they often replicate alongside L1s but use their own distinct replication machinery. In chapter 4, I showed that BovB mobilised SINEs accumulated in gene-poor regions instead of open chromatin gene-rich regions where L1 mobilised SINEs tend to accumulate. Further analysis of such genomes will lead to a more refined perspective on retrotransposon insertion dynamics.

Expanding our analysis to birds would also help provide significant insight into genome evolutionary dynamics of complex organisms. This is because the composition of avian genomes is quite distinct from mammals. For example, bird genomes contain varying numbers of microchromosomes and are significantly smaller than mammalian genomes. Additionally, their genome evolutionary dynamics are comparable to mammals as they have taken place over a similar time-frame. Using DNA gain and loss detection combined with multiple pairwise alignments, it will be possible to identify lineage-specific gain and loss events across the bird phylogeny. The genomic distributions of these gain and loss events can then be compared to the distribution of other types of genomic features. This will help determine if the forces shaping genome size dynamics are constant across both birds and mammals, and may provide insight into the origin and evolutionary trajectories of micro chromosomes.

Collectively, my findings demonstrate that complex mammalian genome architecture is an emergent property of the interactions between a small number of components under various evolutionary constraints.

Appendix A

Supplementary for Chapter 2

Supplementary figures

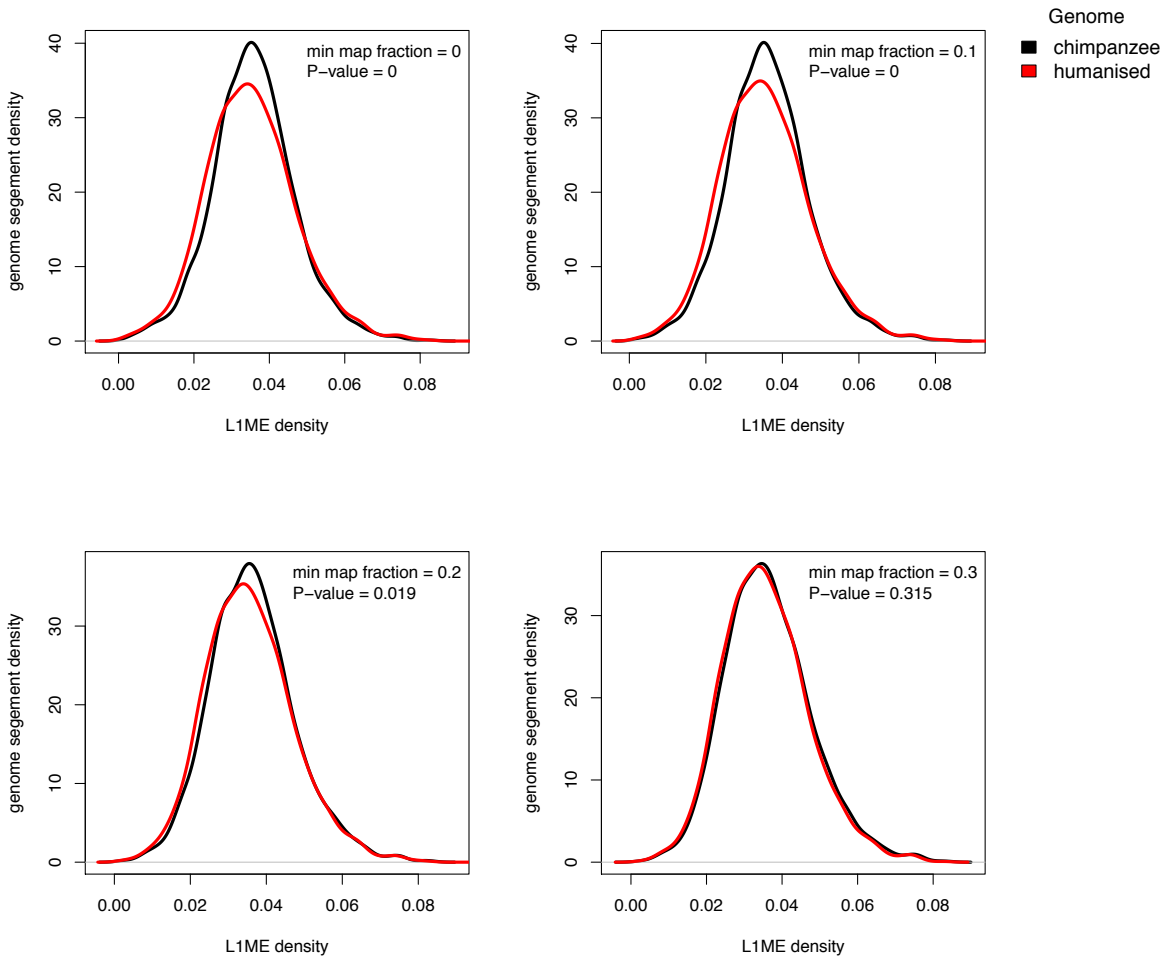


Figure S1: **Minimum mapping fraction.** Increasing the minimum mapping fraction threshold provides humanised retrotransposon genomic distributions a better representation of the remaining proportion of the query species genome. Similarity between distributions is measured by calculating a P-value from the Kolmogorov-Smirnov test. A higher P-value represents a better representation of the unhumanised retrotransposon genomic distribution.

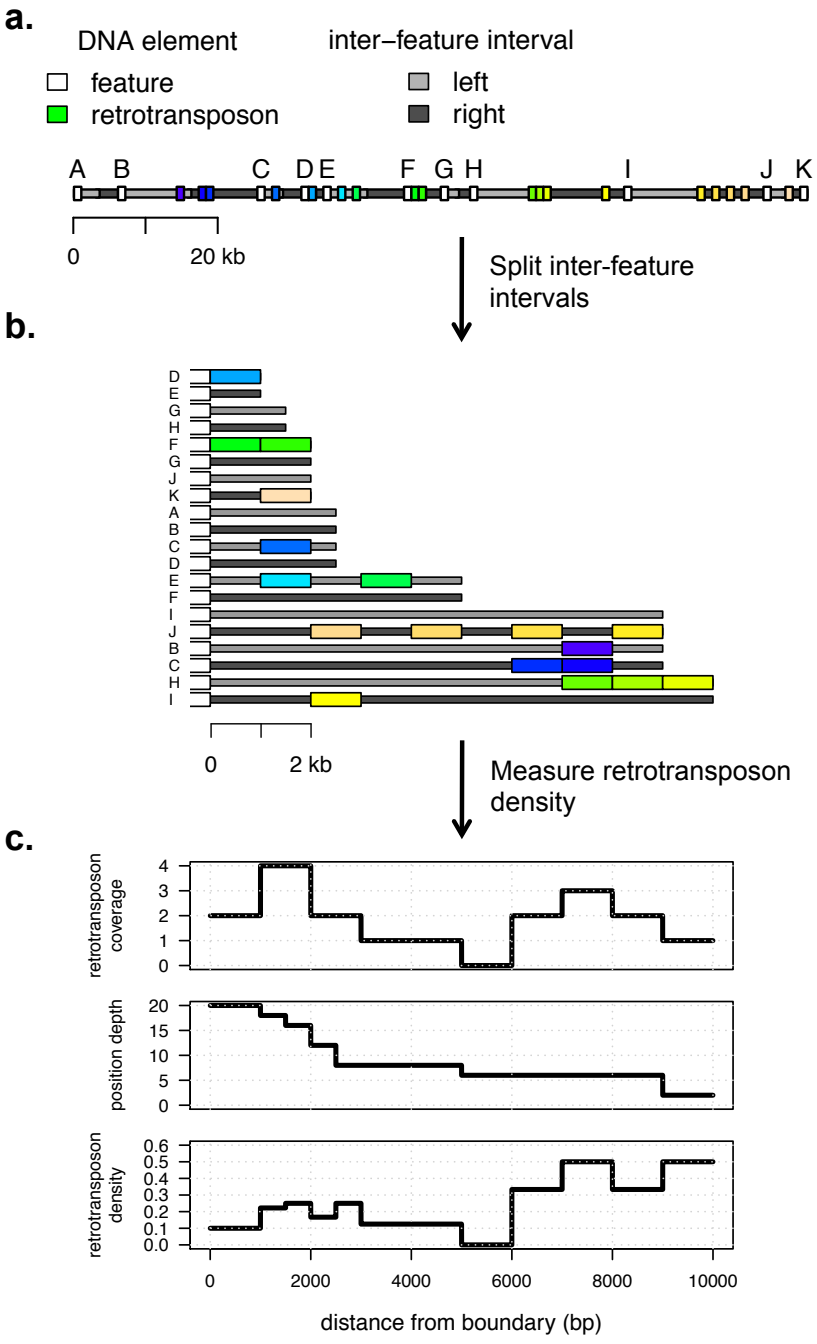


Figure S2: **Example analysis of retrotransposon density at feature boundaries.** **a**, Retrotransposons and features are interspersed across a genome. Feature in this case can refer to one of either gene, exon, or DNase1 cluster. **b**, Inter-feature intervals are split into left and right halves and aligned according to their feature boundaries. **c** From this, the coverage of retrotransposons and the position depth at each position are calculated. Finally, the retrotransposon density at each position is calculated as retrotransposon coverage over position depth.

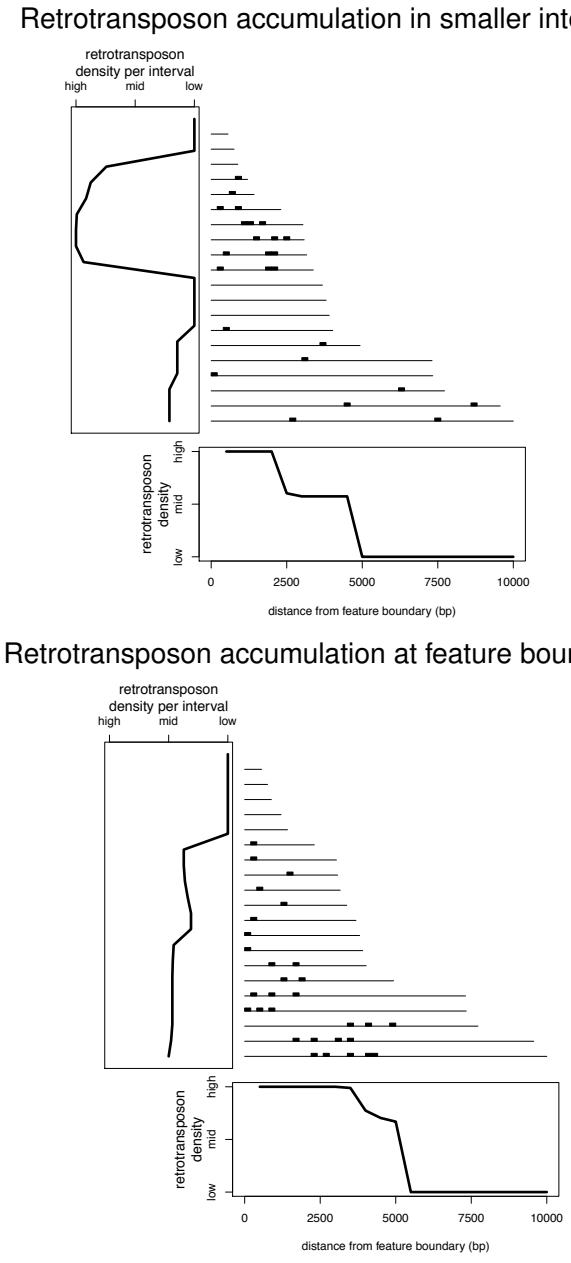


Figure S3: **Example of how interval size accumulation bias can affect retrotransposon density calculations.** Density of retrotransposon insertions across intervals of different size. In both scenarios there appears to be retrotransposon enrichment at feature boundaries. This can result from preferential accumulation into smaller intervals or feature dense regions, rather than preferential retrotransposon accumulation at feature boundaries.

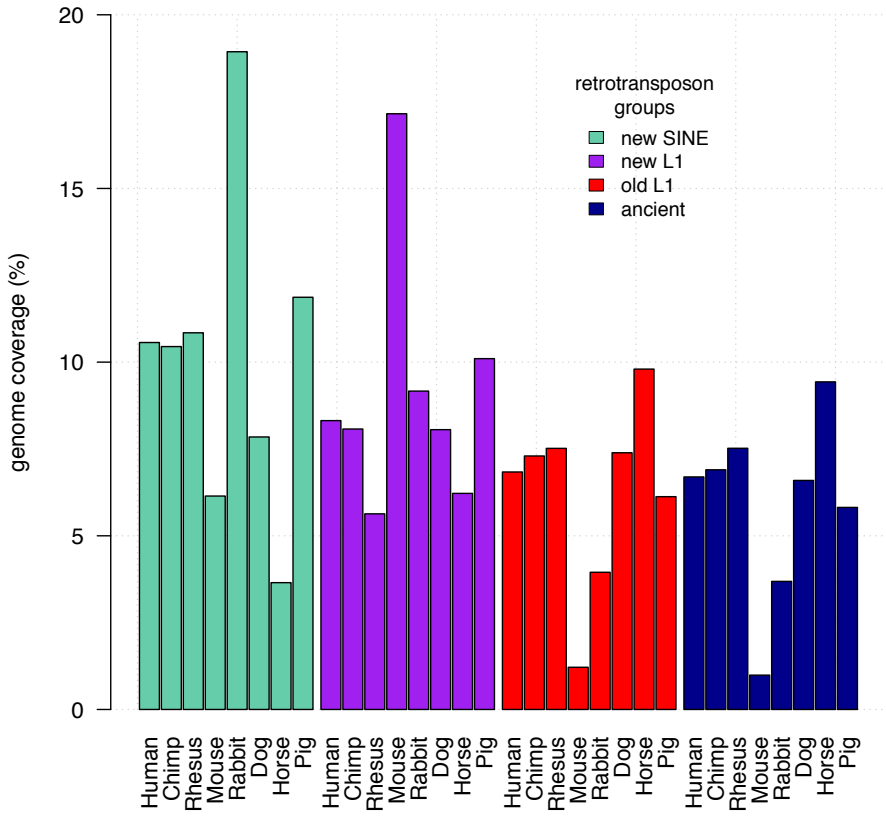


Figure S4: **Percent genome coverage of retrotransposon groups.**

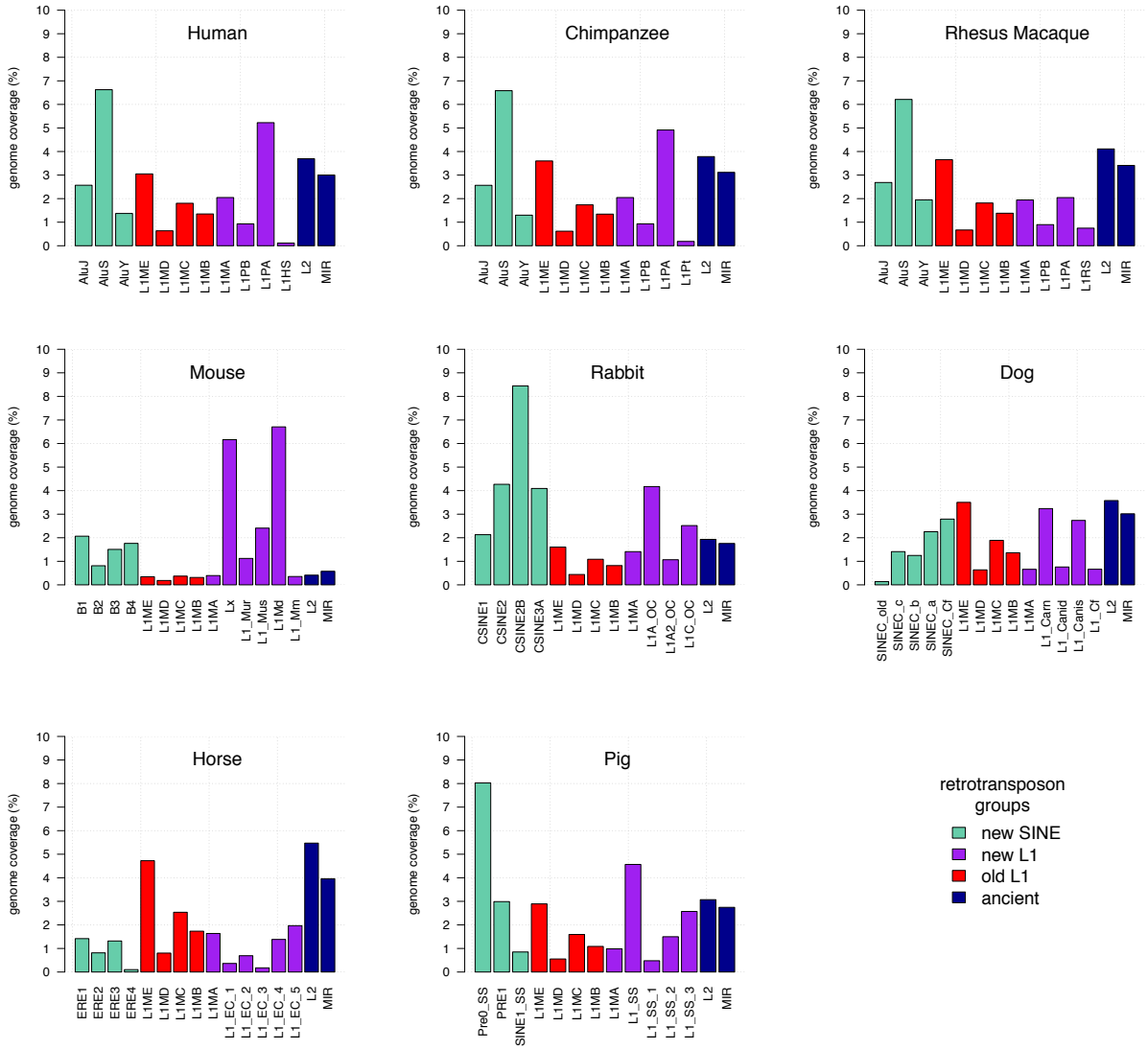


Figure S5: **Percent genome coverage of retrotransposon families.**

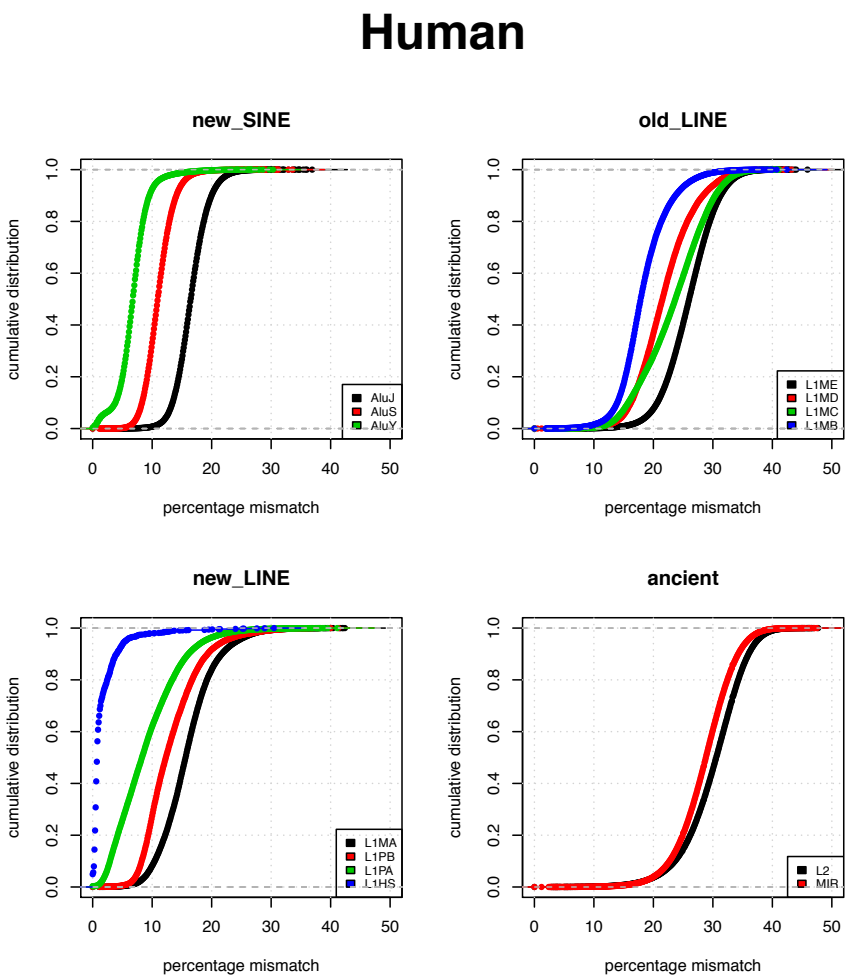


Figure S6: **Human mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

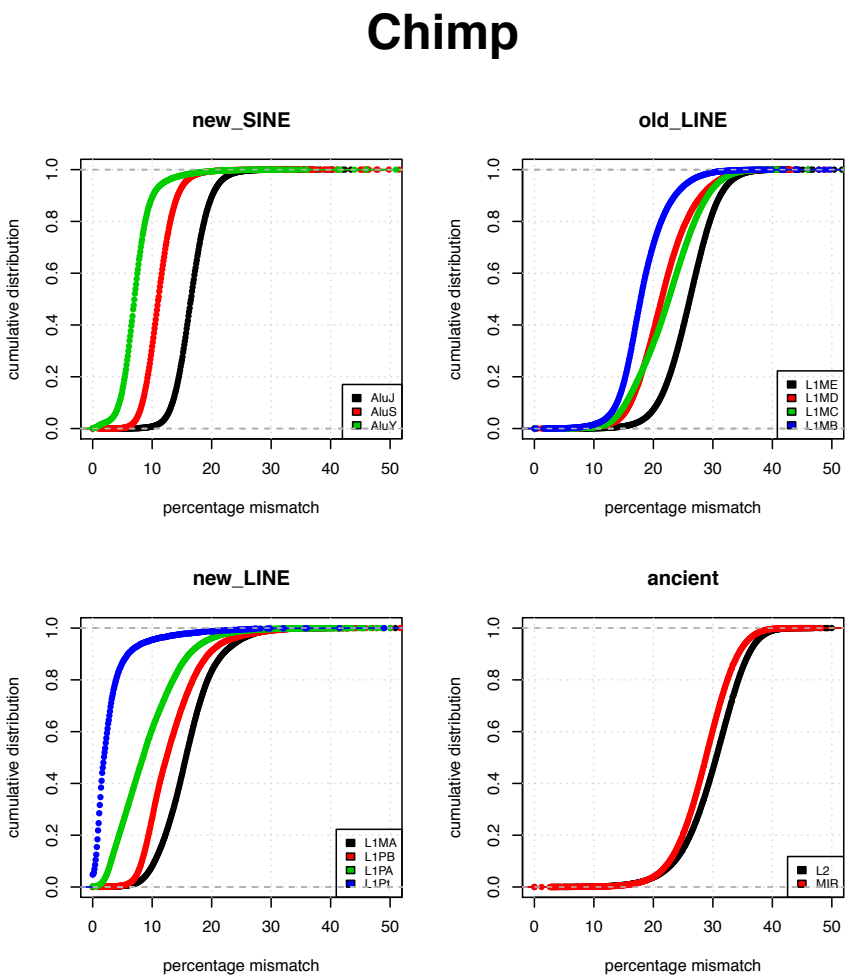


Figure S7: **Chimpanzee mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

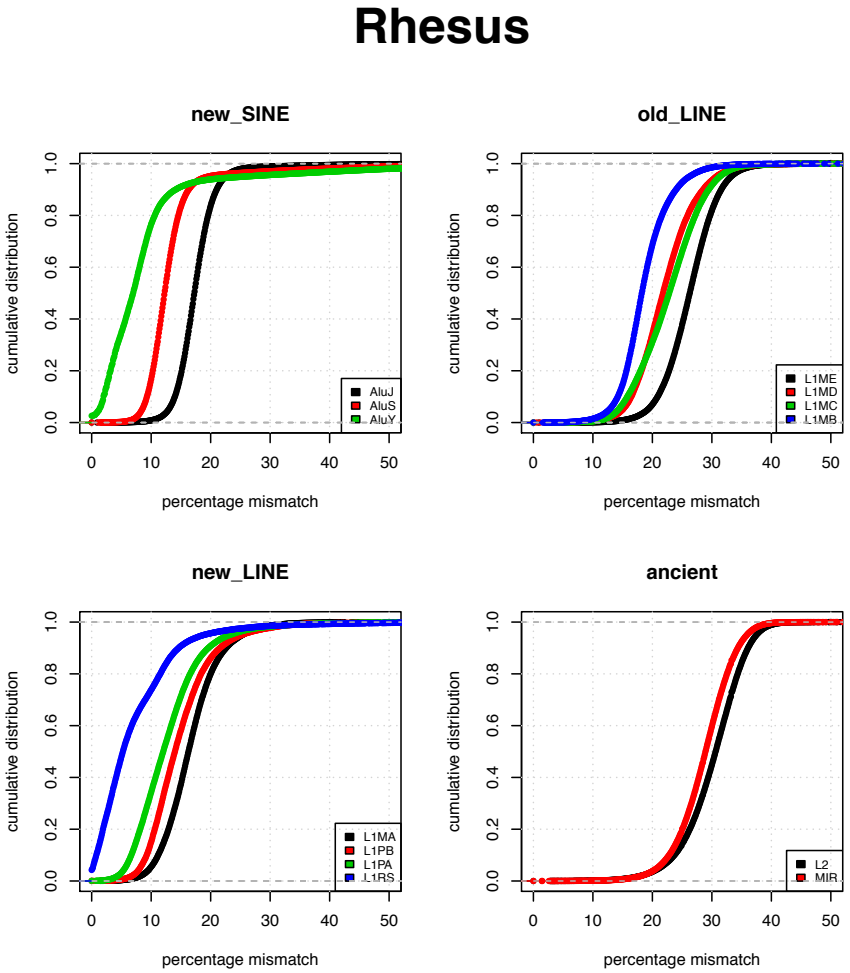


Figure S8: **Rhesus Macaque mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

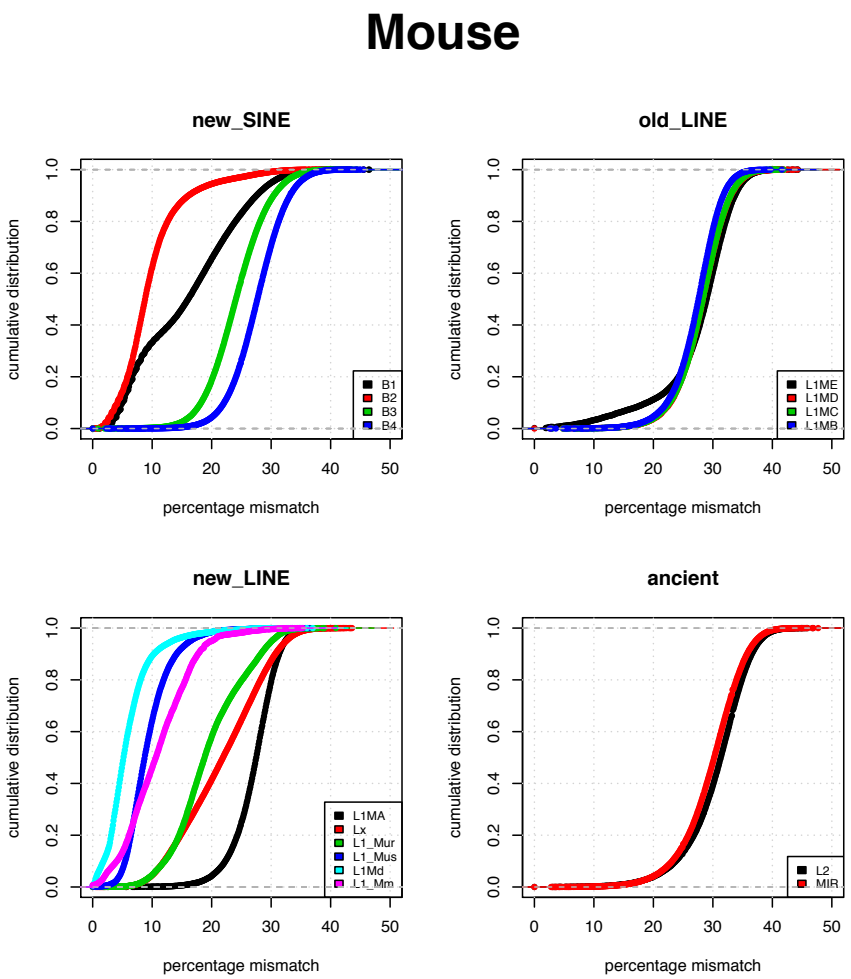


Figure S9: **Mouse mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

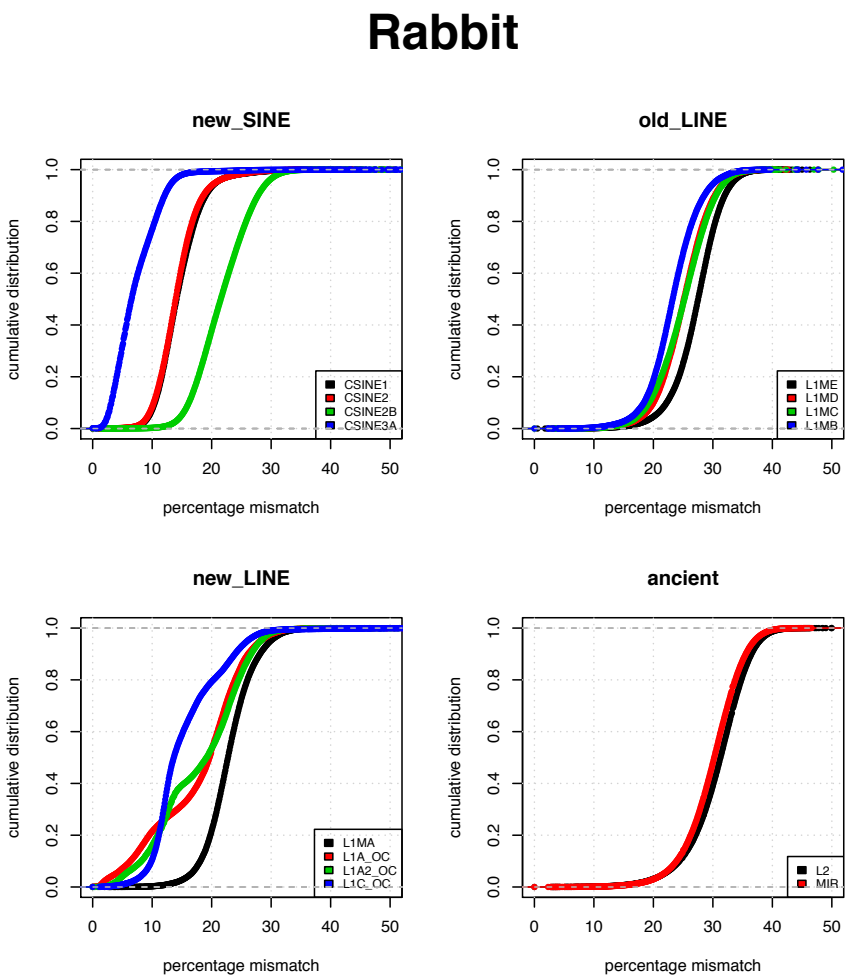


Figure S10: **Rabbit mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

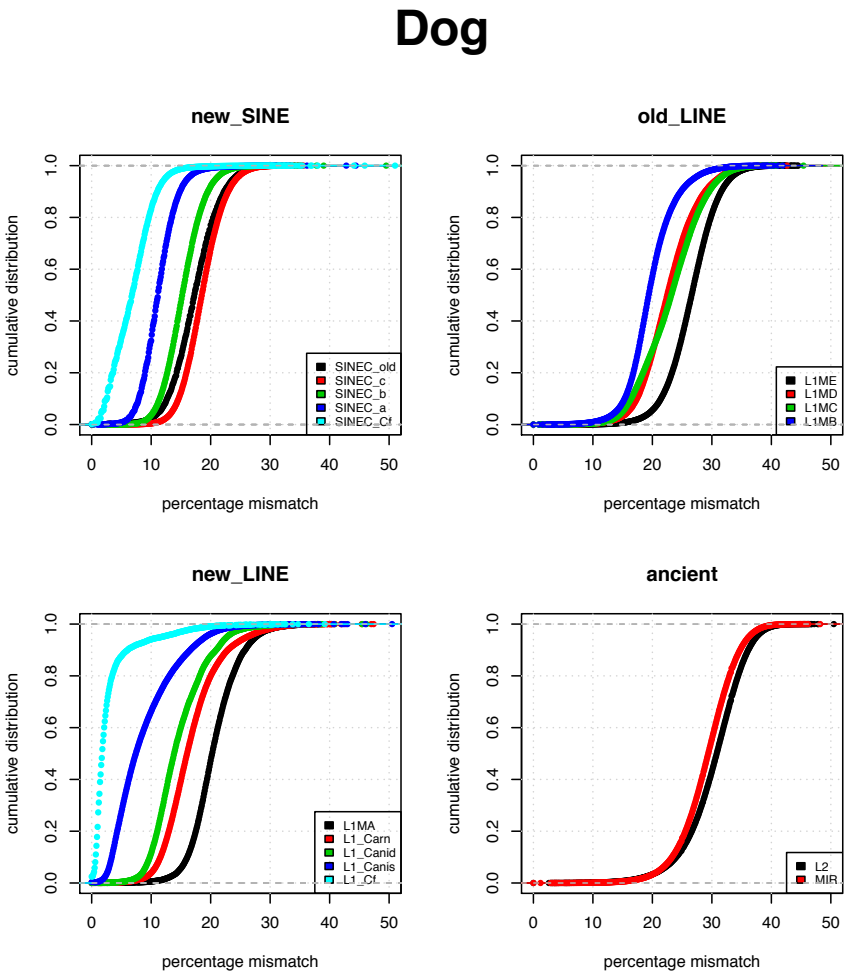


Figure S11: **Dog mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

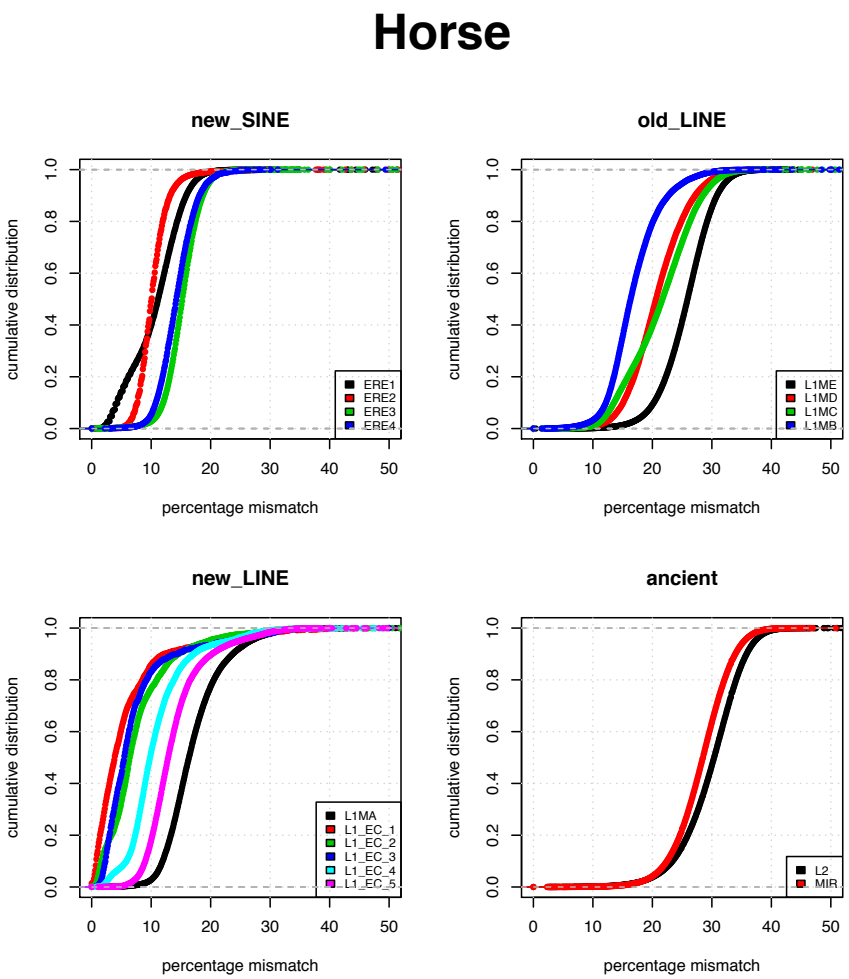


Figure S12: **Horse mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.

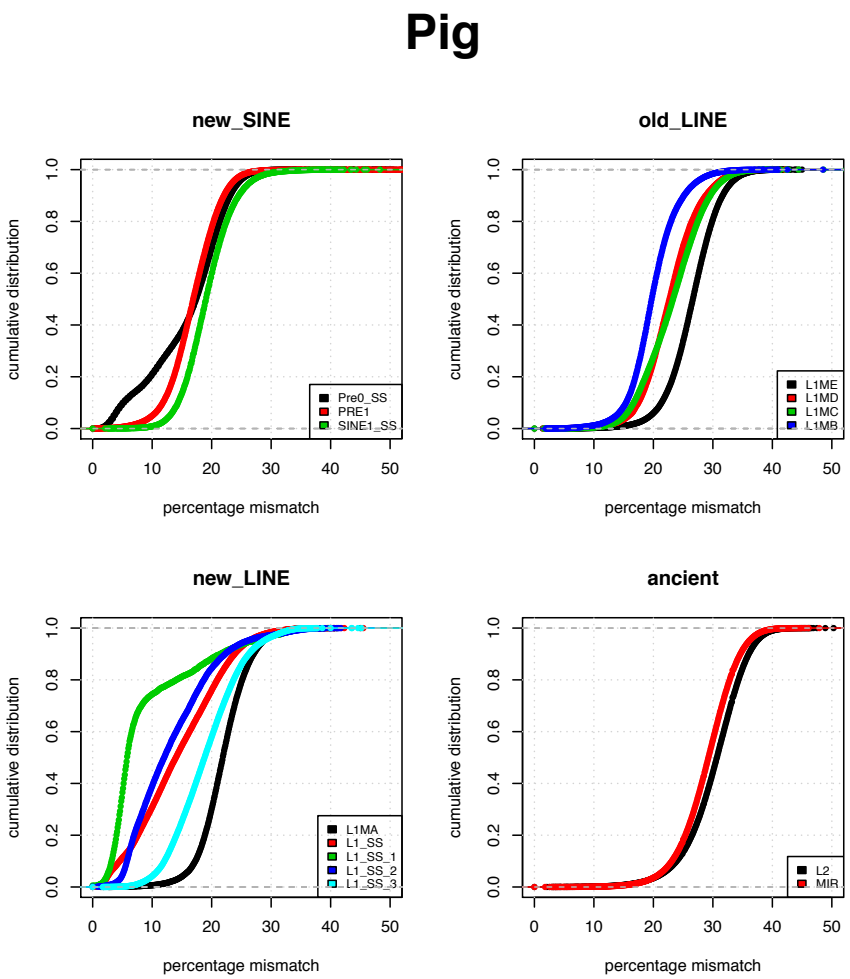
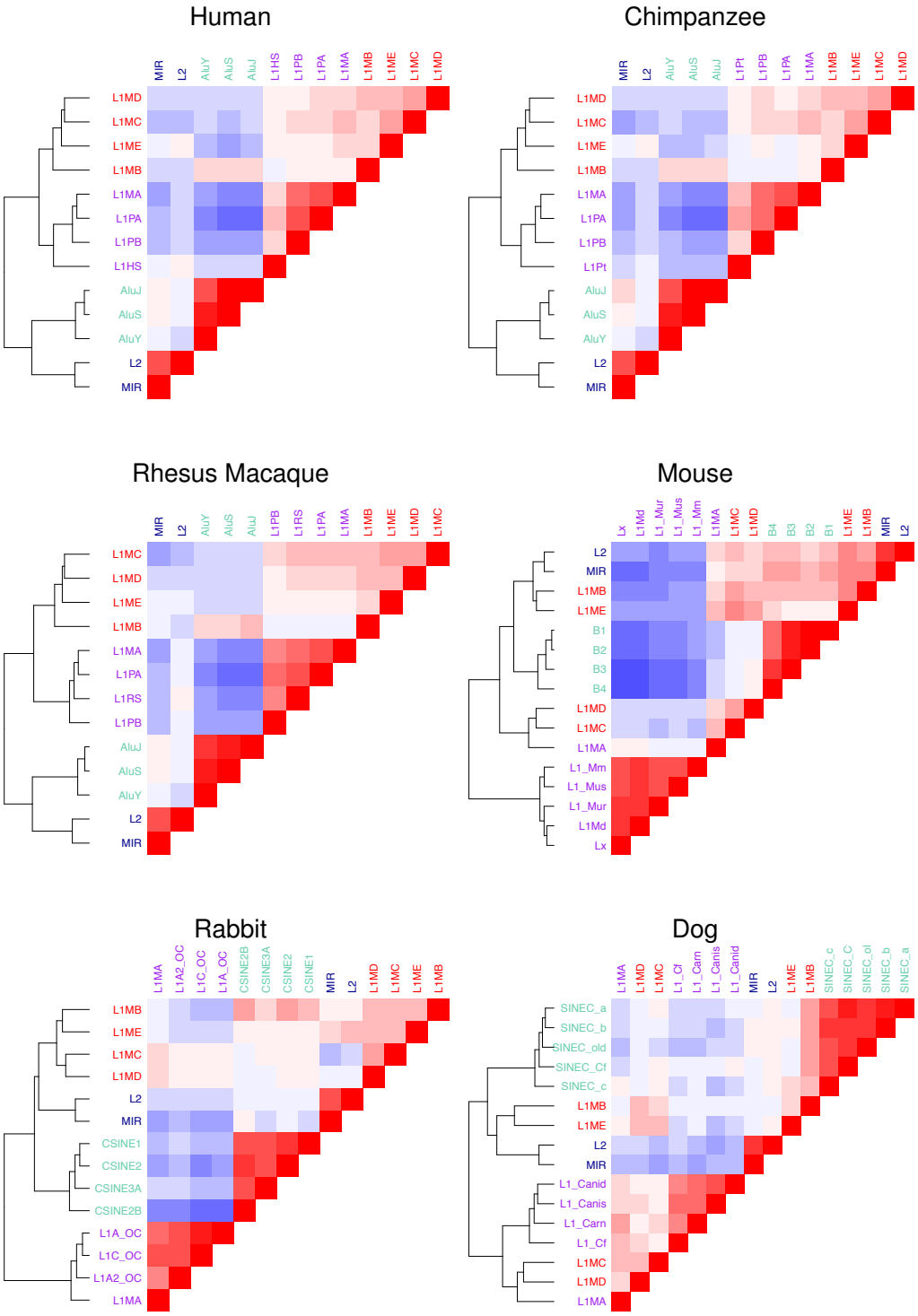


Figure S13: **Pig mismatch scores for each retrotransposon family.** Scores are calculated based on percentage mismatch of each repeat element from consensus sequence.


Human
MIR L2 AluY AluS AluJ L1HS L1PB L1PA L1MA L1MB L1ME L1MC L1MD
L1MD L1MC L1ME L1MB L1MA L1PA L1PB L1HS AluJ AluS AluY L2 MIR
Chimpanzee
MIR L2 AluY AluS AluJ L1Pt L1PB L1PA L1MA L1MB L1ME L1MC L1MD
L1MD L1MC L1ME L1MB L1MA L1PA L1PB L1Pt AluJ AluS AluY L2 MIR
Rhesus Macaque
MIR L2 AluY AluS AluJ L1PB L1RS L1PA L1MA L1MB L1ME L1MD L1MC
L1MC L1MD L1ME L1MB L1MA L1PA L1RS L1PB AluJ AluS AluY L2 MIR
Mouse
Lx L1Md L1_Mur L1_Mus L1_Mm L1MA L1MC L1MD B4 B3 B2 B1 L1ME L1MB MIR L2
L2 MIR L1MB L1ME B1 B2 B3 B4 L1MD L1MC L1MA L1_Mm L1_Mus L1_Mur L1Md Lx
Rabbit
L1MA L1A2_OC L1C_OC L1A_OC CSINE2B CSINE3A CSINE2 CSINE1 MIR L2 L1MD L1MC L1ME L1MB
L1MB L1ME L1MC L1MD L2 MIR CSINE1 CSINE2 CSINE3A CSINE2B L1A_OC L1C_OC L1A2_OC L1MA
Dog
L1MA L1MD L1MC L1_Cf L1_Carn L1_Canis L1_Canid MIR L2 L1ME L1MB SINEC_c SINEC_C SINEC_ol SINEC_b SINEC_a
SINEC_a SINEC_b SINEC_old SINEC_Cf SINEC_c L1MB L1ME L2 MIR L1_Canid L1_Canis L1_Carn L1_Cf L1MC L1MD L1MA

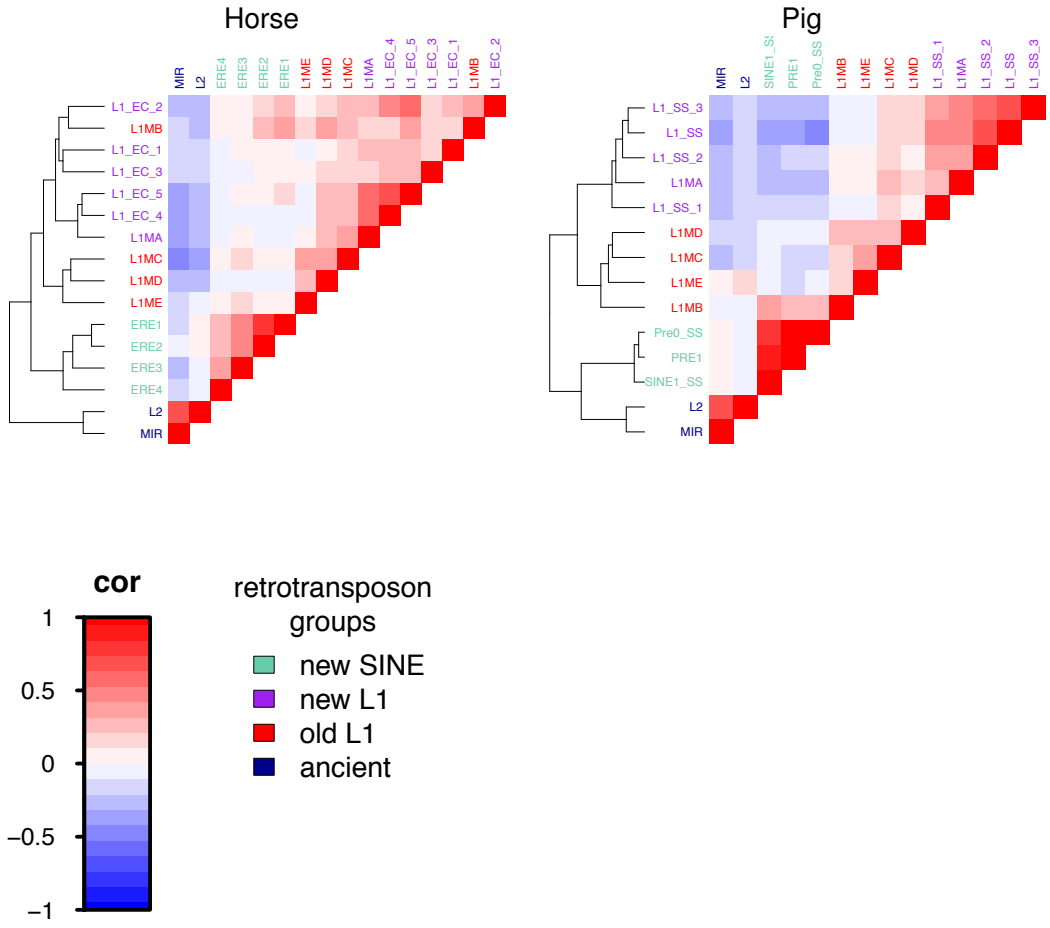


Figure S14: **Pairwise genomic correlations of retrotransposon distributions.** Pairwise correlation analysis was carried out on retrotransposon family densities in 1 Mb genomic segments.

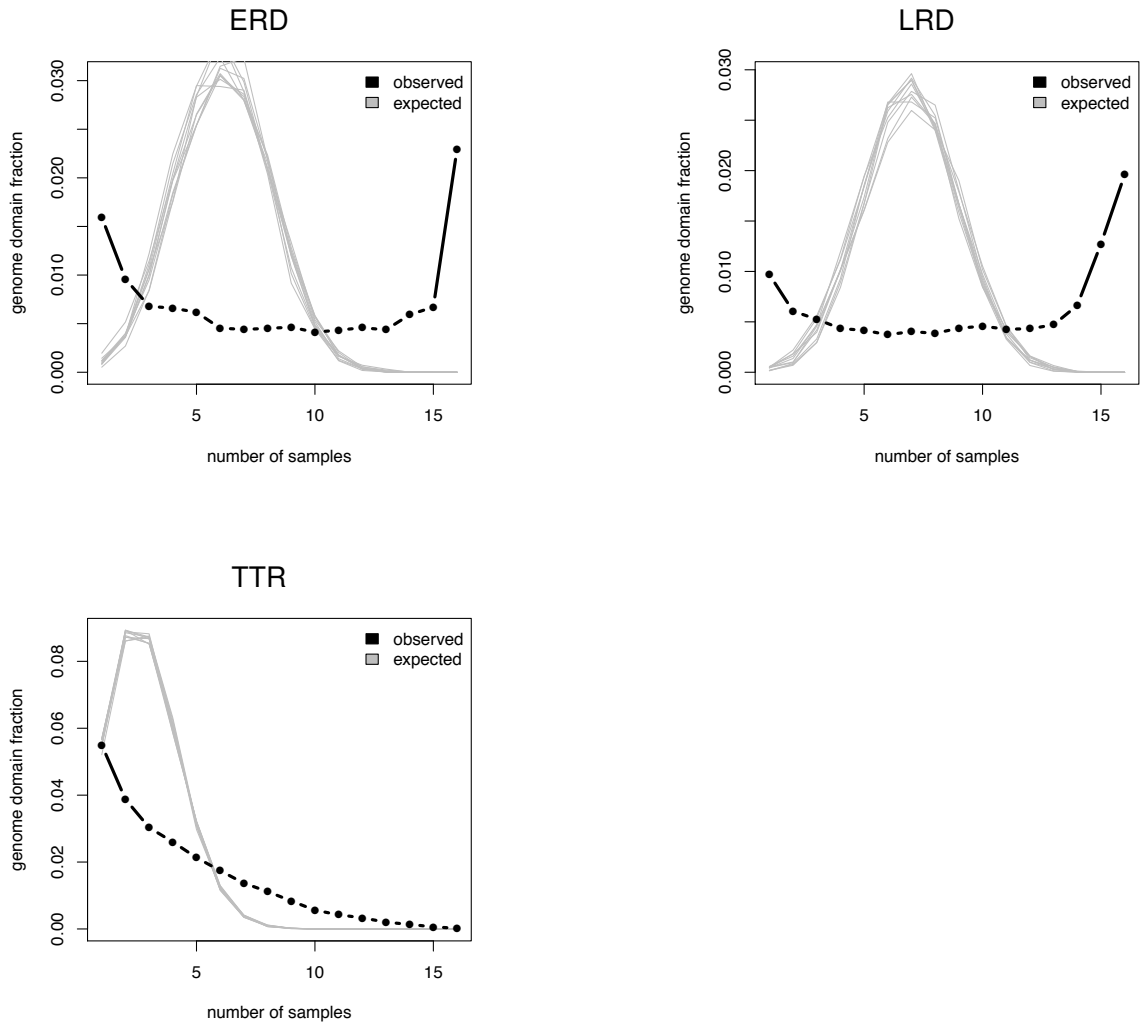


Figure S15: **Overlapping fractions of each RD type.** The overlapping fraction of a particular RD type across 16 RD samples. To calculate our expected overlapping fraction, we randomly positioned each RD type from each sample. This process was repeated 10 times.

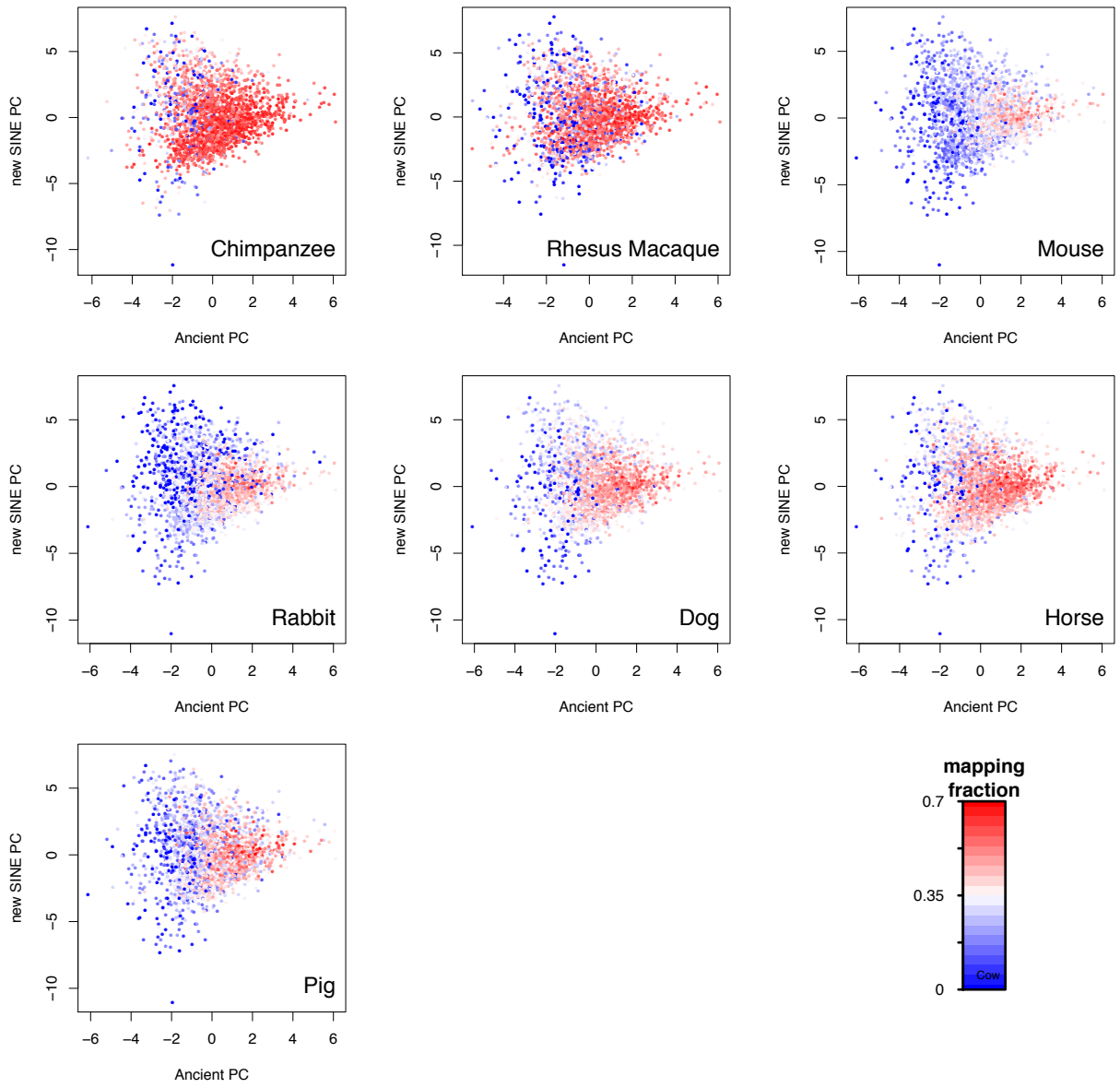


Figure S16: **Mapping fractions of the human genome.** PCA biplots of human segments, where each segment is coloured according to the fraction that maps to a segment in a given non-human query species.

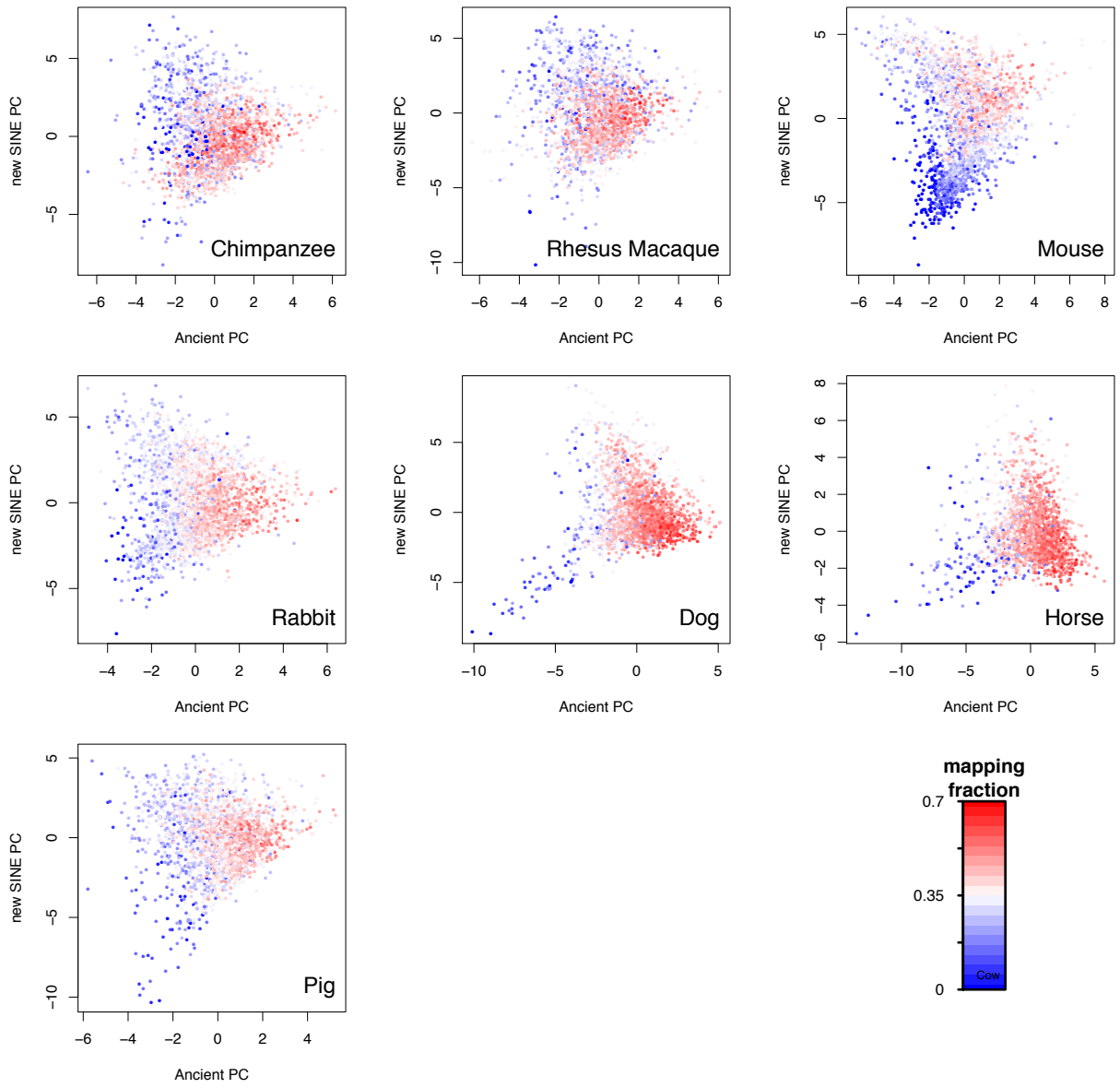


Figure S17: **Mapping fractions of each non-human speceis' genome.** PCA biplots of genome segments from each non-human query species, where each segment is coloured according to the fraction that maps to human segments.

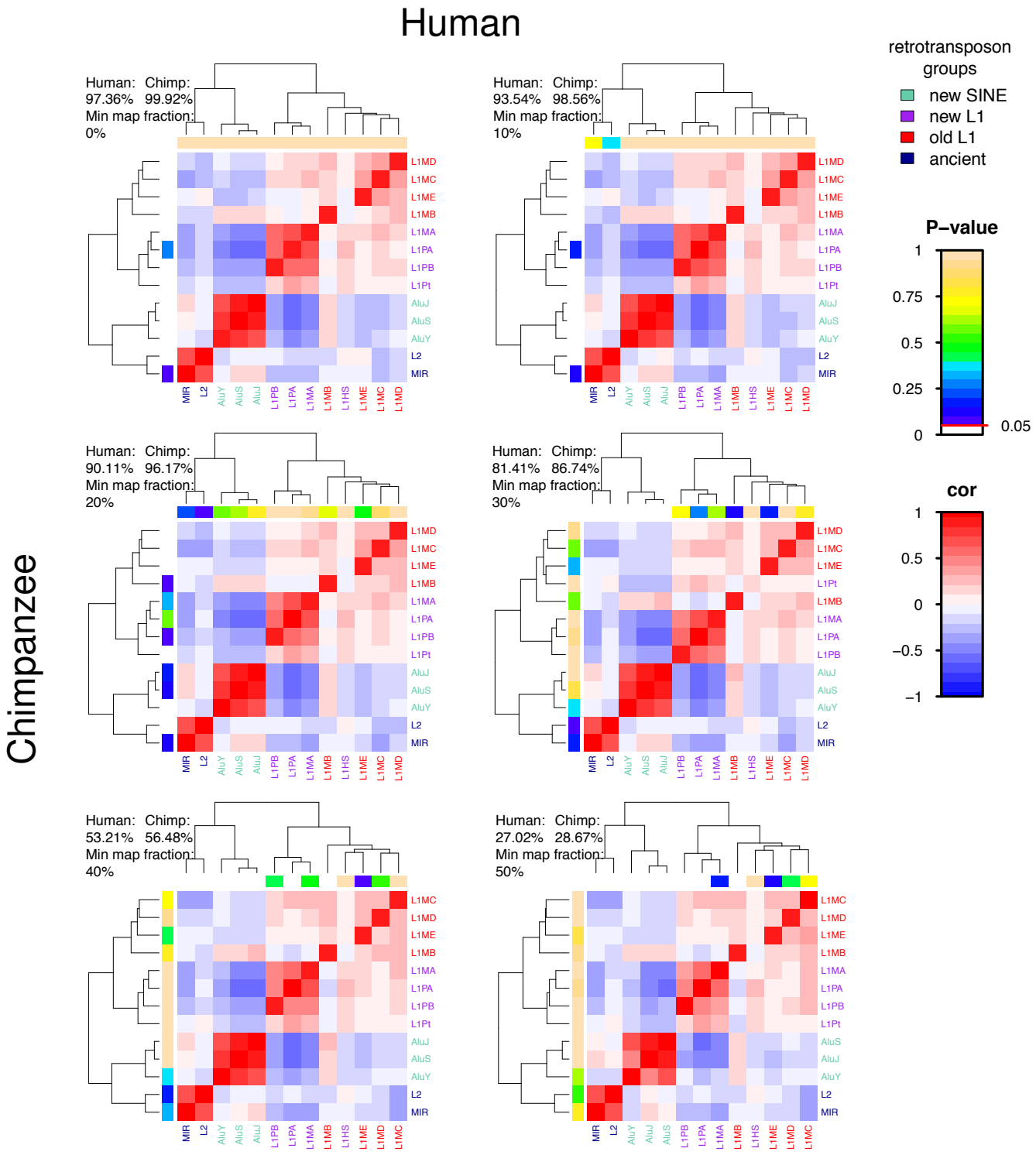


Figure S18: **Pairwise correlation comparisons of human and humanised chimpanzee retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Chimpanzee P-values represent the effect of humanising on the filtered chimpanzee retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

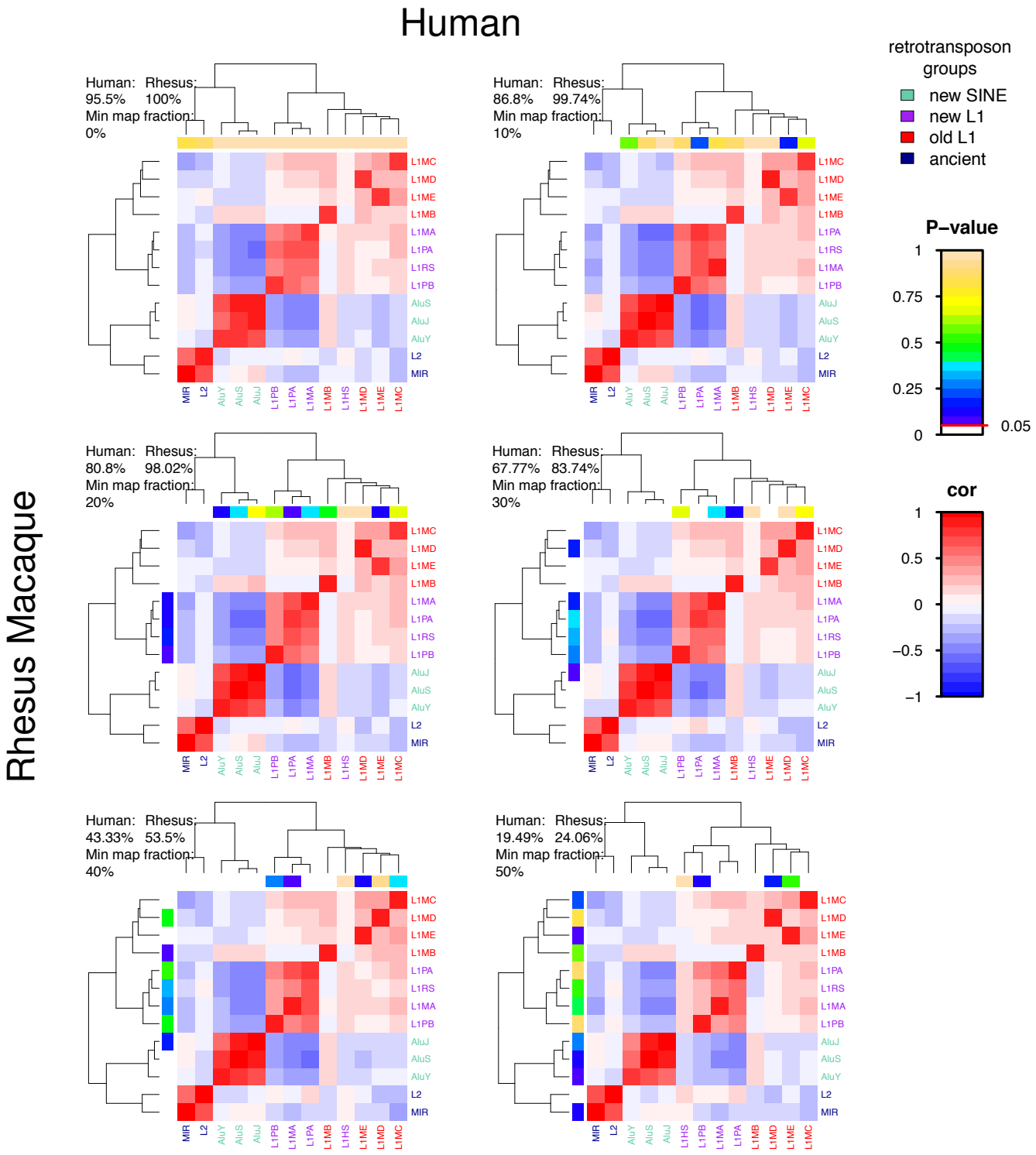


Figure S19: **Pairwise correlation comparisons of human and humanised rhesus macaque retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Rhesus macaque P-values represent the effect of humanising on the filtered rhesus macaque retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

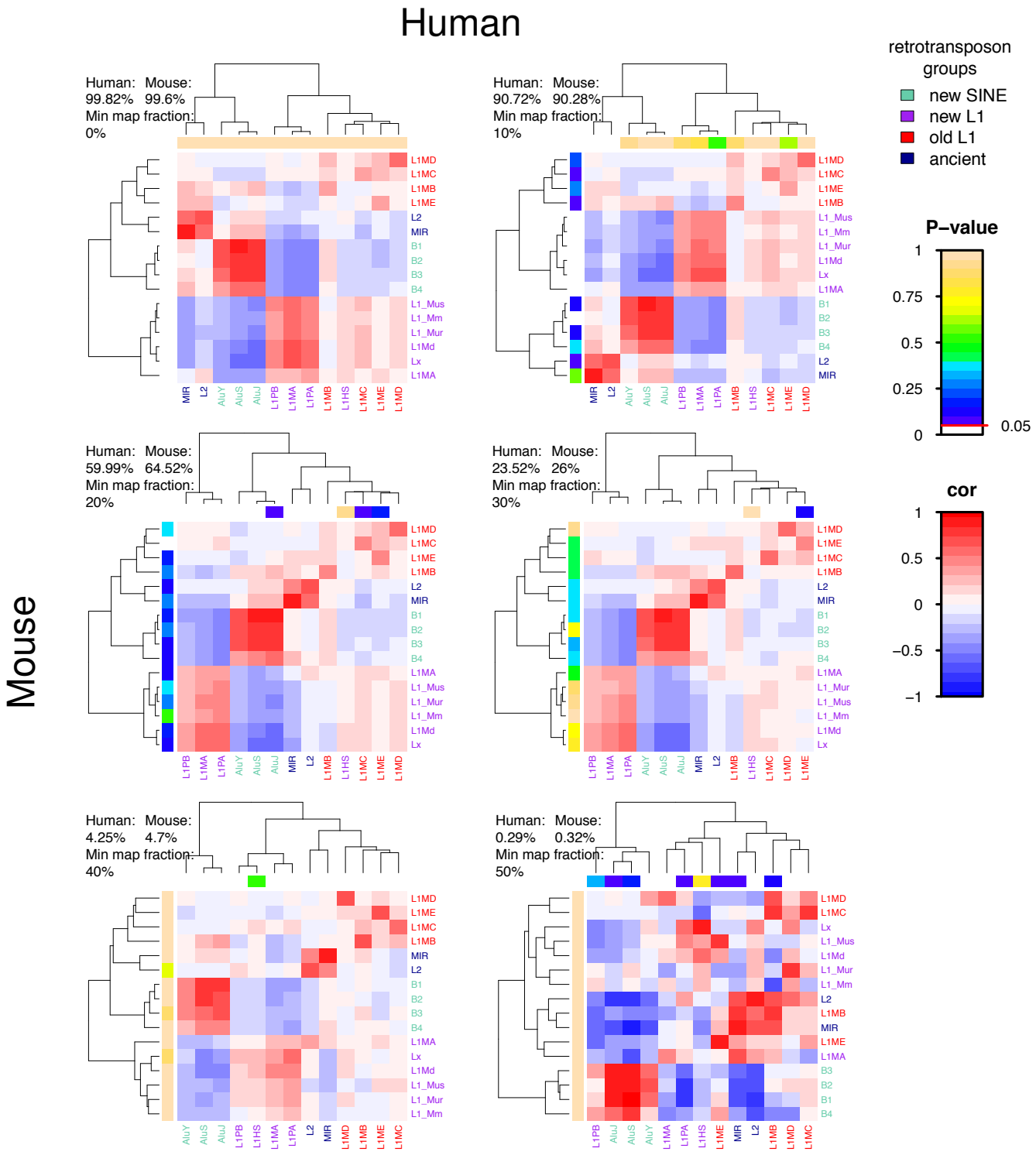


Figure S20: **Pairwise correlation comparisons of human and humanised mouse retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Mouse P-values represent the effect of humanising on the filtered mouse retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

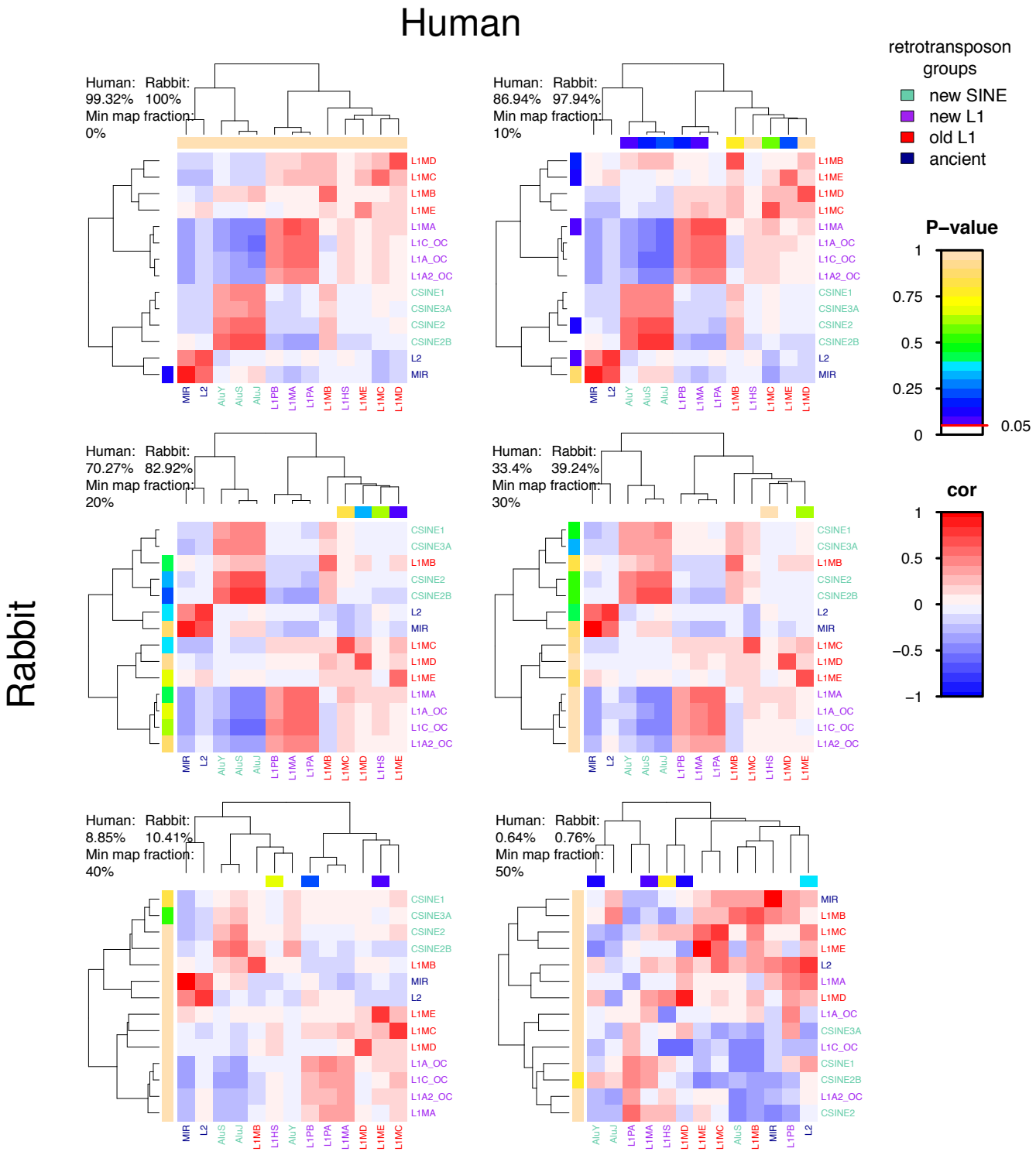


Figure S21: **Pairwise correlation comparisons of human and humanised rabbit retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Rabbit P-values represent the effect of humanising on the filtered rabbit retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

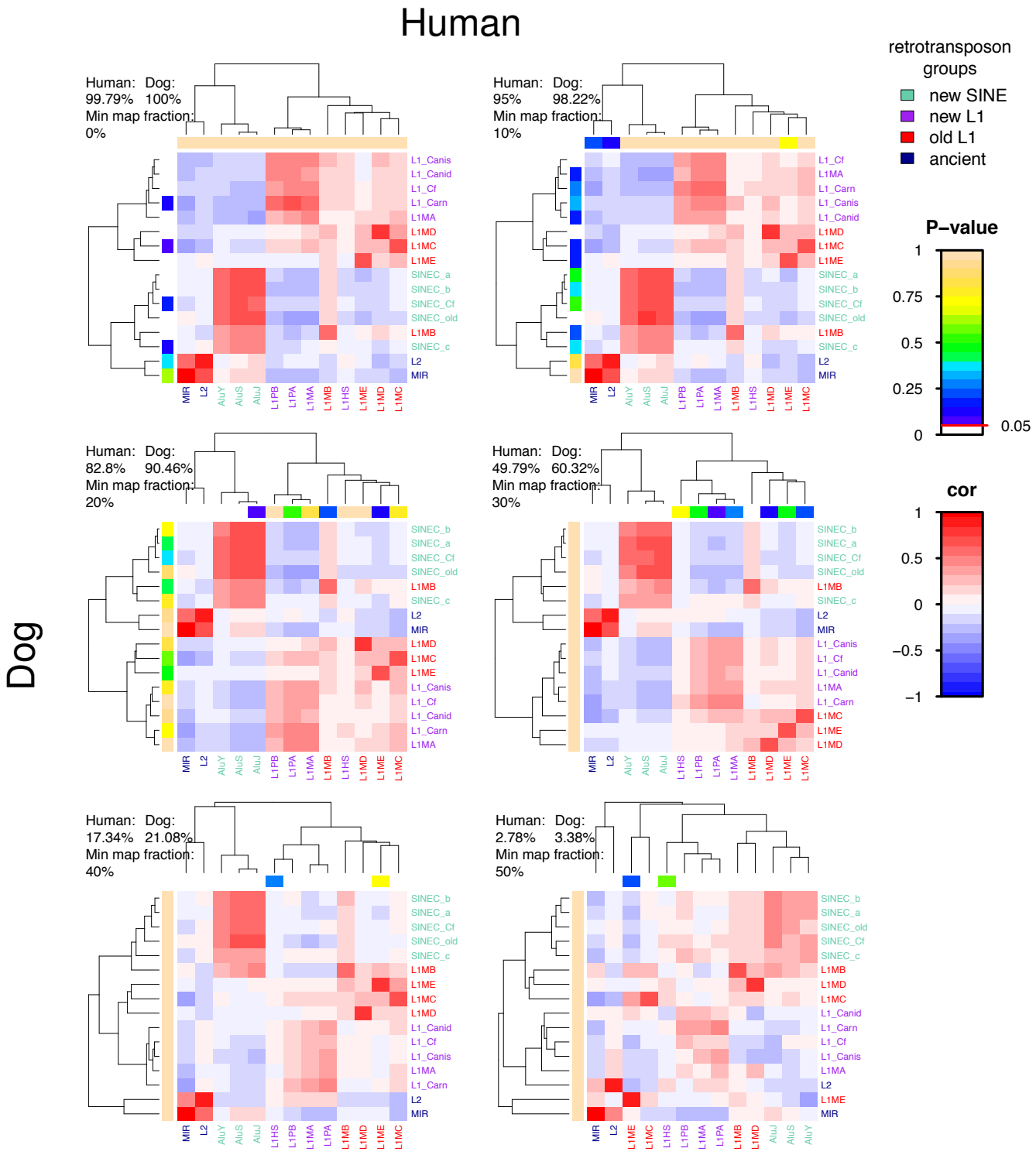


Figure S22: **Pairwise correlation comparisons of human and humanised dog retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Dog P-values represent the effect of humanising on the filtered dog retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

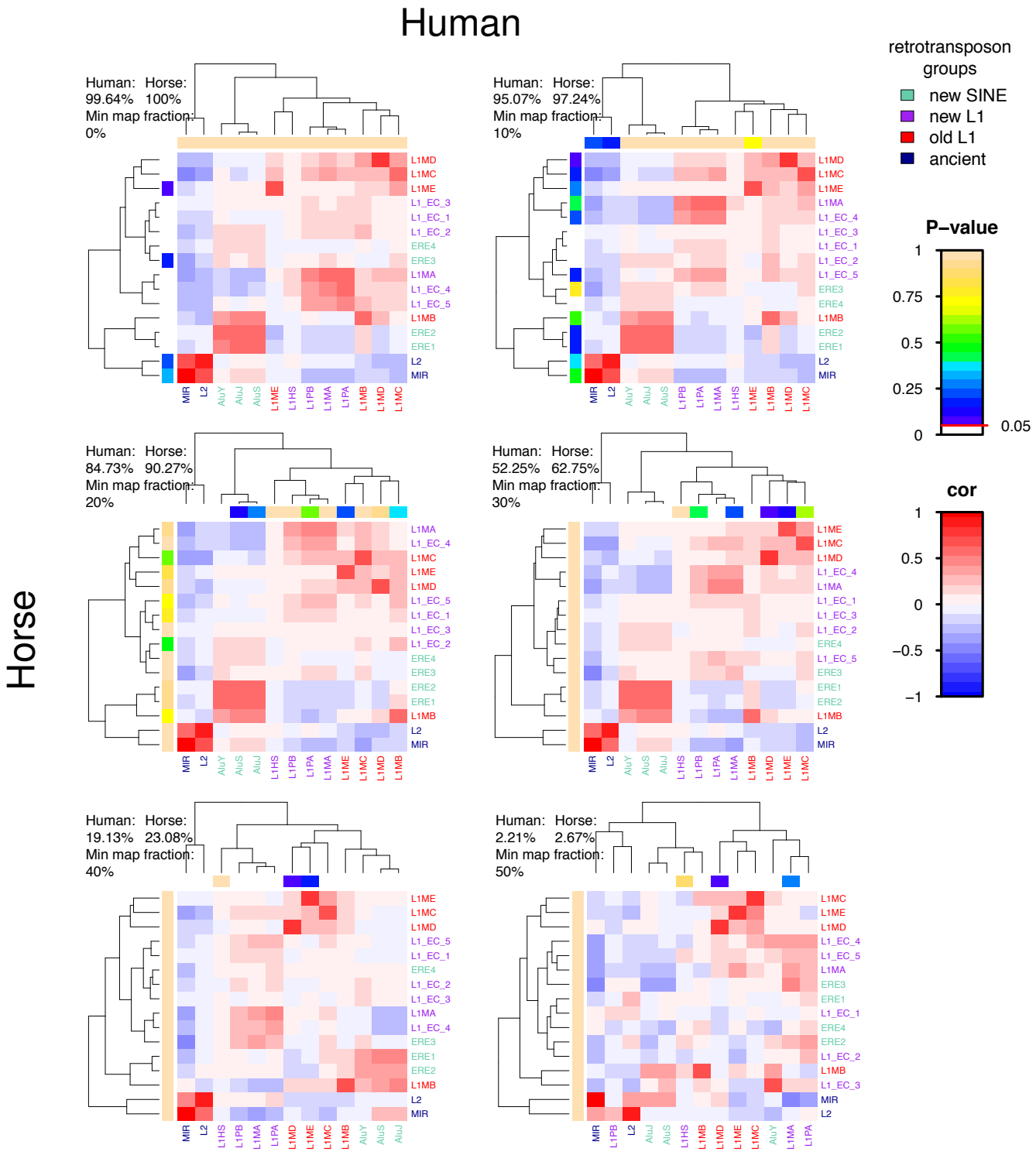


Figure S23: **Pairwise correlation comparisons of human and humanised horse retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Horse P-values represent the effect of humanising on the filtered horse retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

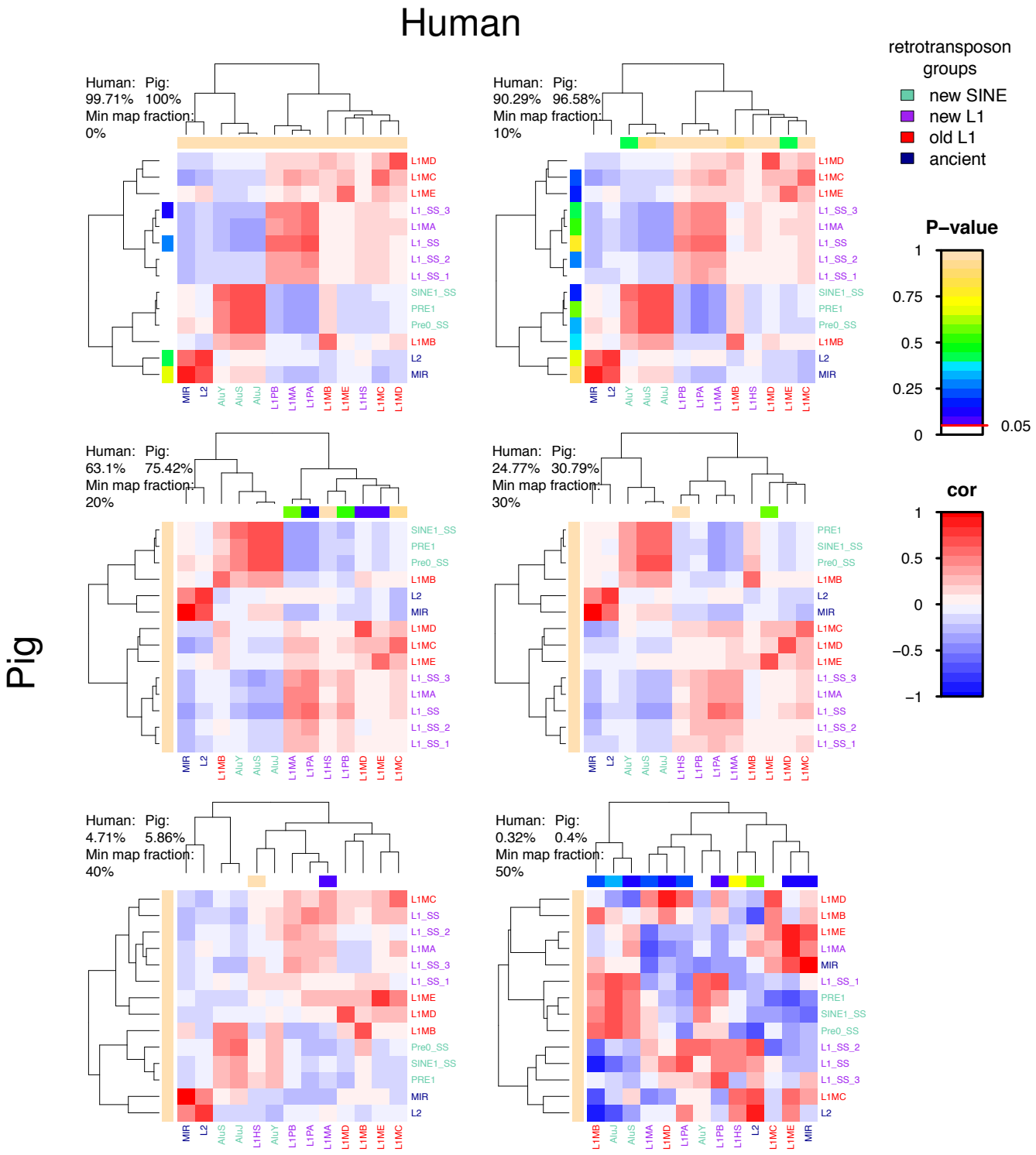


Figure S24: **Pairwise correlation comparisons of human and humanised pig retrotransposon genomic distributions.** Values in the top left reflect the proportion of each genome analysed after filtering at each minimum mapping fraction threshold. Heatmap colours represent Pearson's correlation coefficient. Pig P-values represent the effect of humanising on the filtered pig retrotransposon density distribution. Human P-values represent the effect of filtering on the human retrotransposon density distribution.

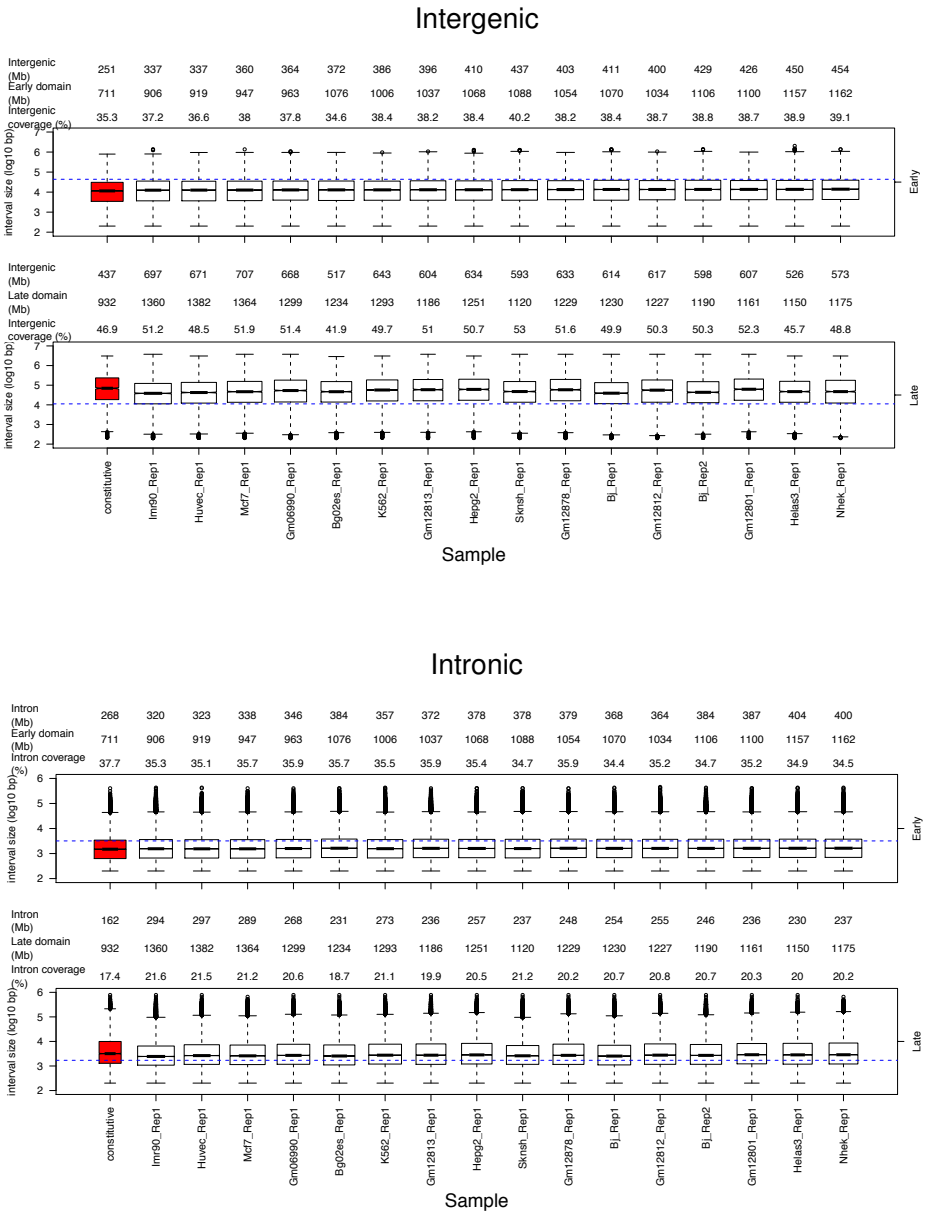


Figure S25: **Genome properties of replication domains across each dataset.** Constitutive replication timing domains were identified for both ERDs and LRDs in intergenic and intronic regions. Blue dotted lines represent the mean interval size for intervals found in regions of opposite replication timing. Constitutive domains are coloured in red. The width of each box represents the relative number of intergenic or intronic regions per cell line in either ERDs or LRDs. Cell lines are ordered by mean interval size of intergenic regions in ERDs from lowest to highest.

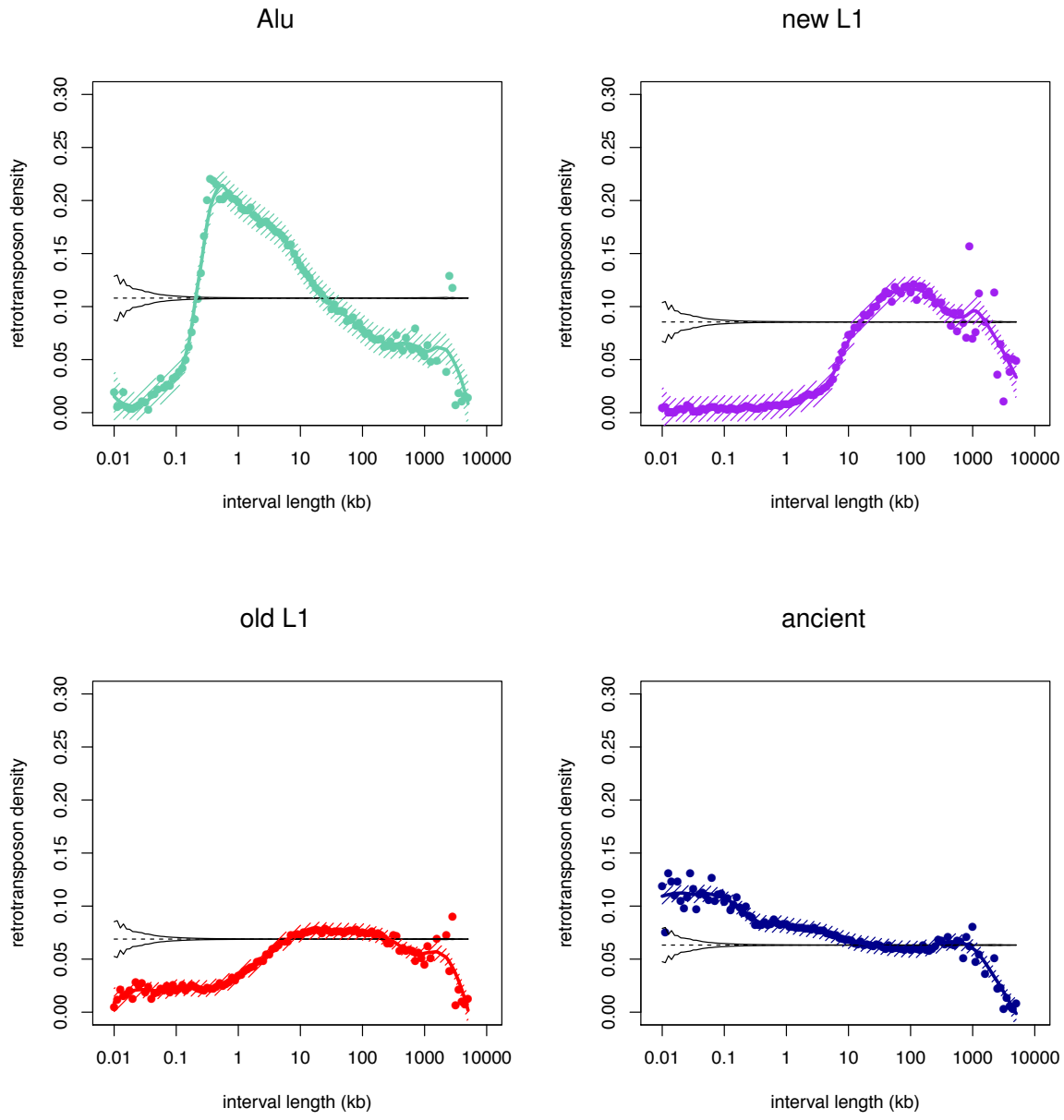


Figure S26: **Interval size bias of retrotransposon density for intervals between DNase1 clusters.** Retrotransposon density for different sized intervals is represented at each point. A line was fitted using LOESS and standard error estimates are represented by cross hatching. The expected retrotransposon density at each interval size is shown by the dotted line. A confidence interval of 3 standard deviations from expected is represented by black solid lines.

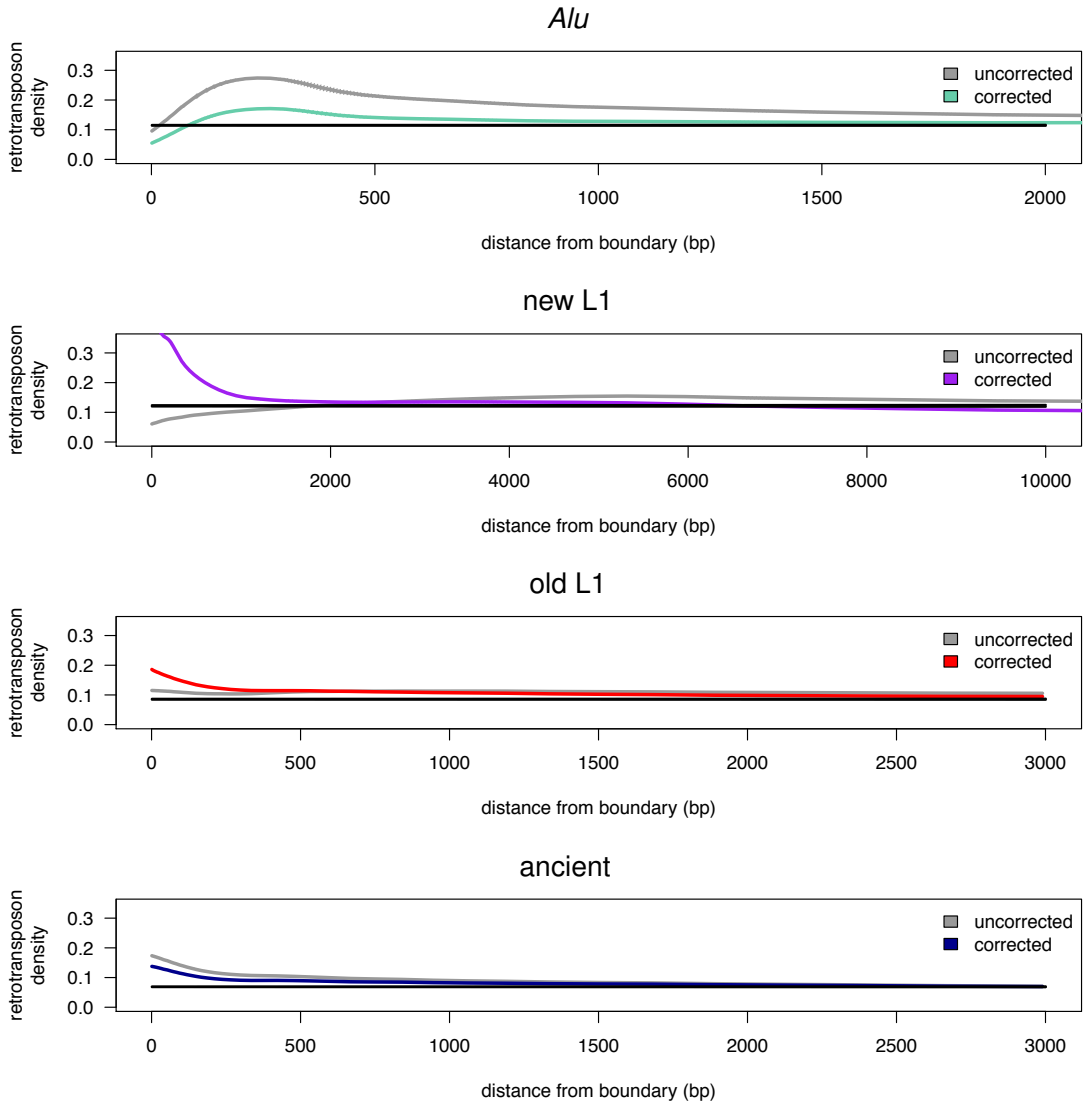


Figure S27: **Smoothed corrected retrotransposon density at the boundaries of DNase1 clusters.** Retrotransposon densities were corrected for bias levels corresponding to the interval sizes present at each position from the boundary. A confidence interval of 3 standard deviations from expected is represented by black solid lines Retrotransposon densities were smoothed by using an expanding window from which variance was calculated. Retrotransposon density variance is represented by cross hatching at three standard deviations. However, in this case levels of variance for both expected and smoothed retrotransposon densities are negligible.

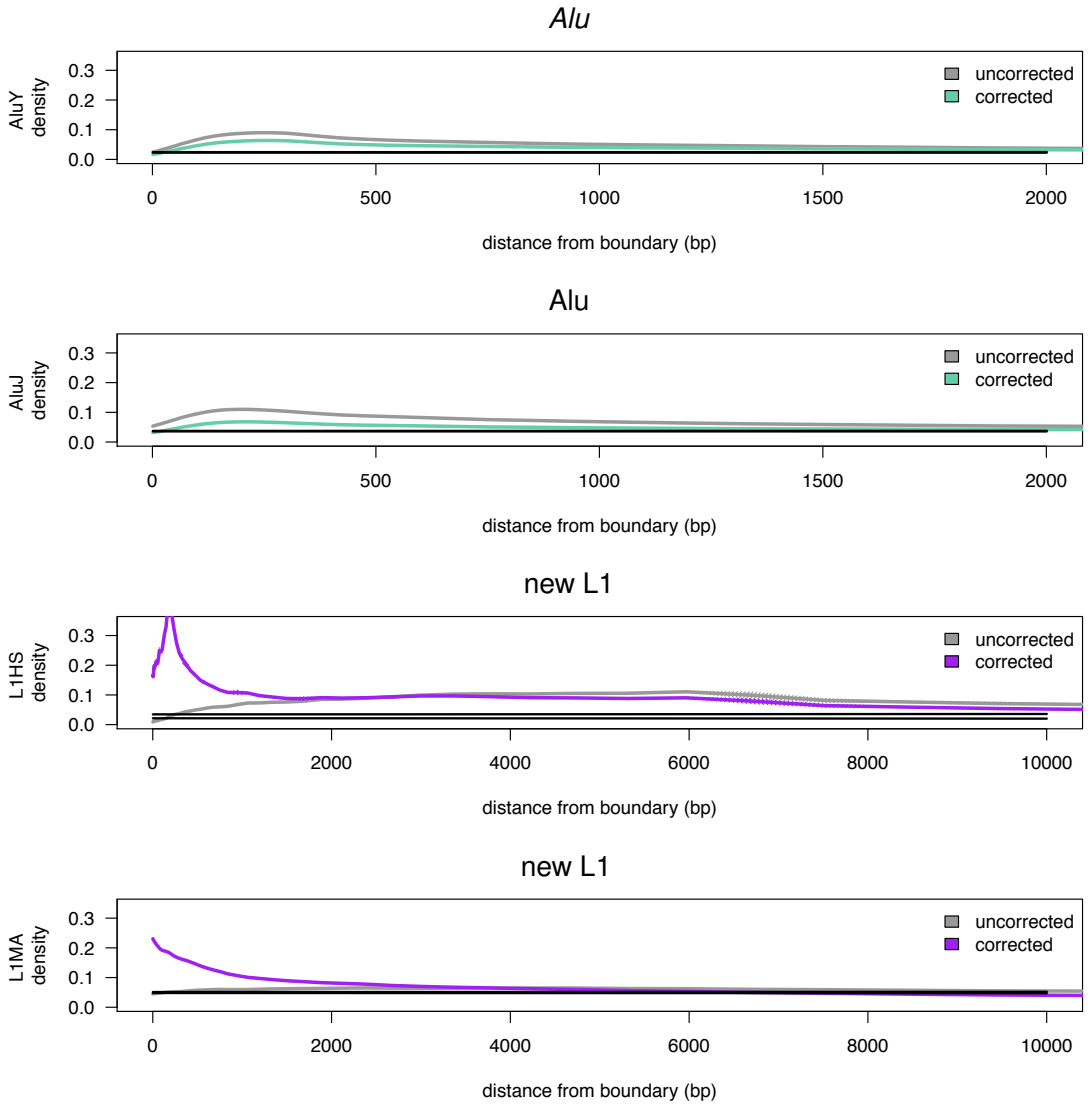


Figure S28: **Smoothed corrected Alu and new L1 density at the boundaries of DNase1 clusters.** Retrotransposon densities were corrected for bias levels corresponding to the interval sizes present at each position from the boundary. A confidence interval of 3 standard deviations from expected is represented by black solid lines Retrotransposon densities were smoothed by using an expanding window from which variance was calculated. Retrotransposon density variance is represented by cross hatching at three standard deviations. However, in this case levels of variance for both expected and smoothed retrotransposon densities are negligible. For both Alu and new L1 the most recent and least recent families are displayed.

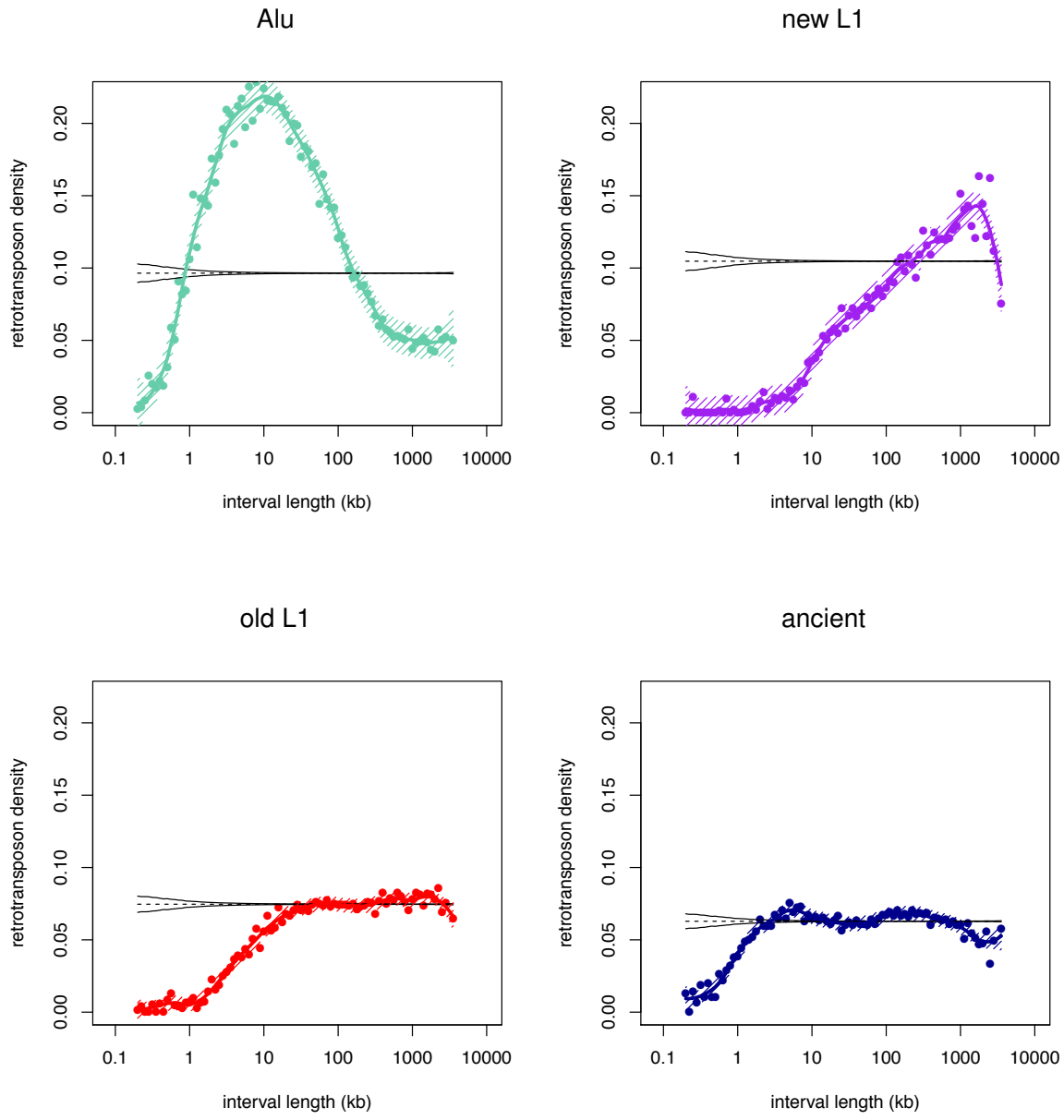


Figure S29: **Interval size bias of retrotransposon density for intergenic regions.** Retrotransposon density for different sized intervals is represented at each point. A line was fitted using LOESS and standard error estimates are represented by cross hatching. The expected retrotransposon density at each interval size is represented by the dotted line. A confidence interval of 3 standard deviations from expected is represented by black solid lines.

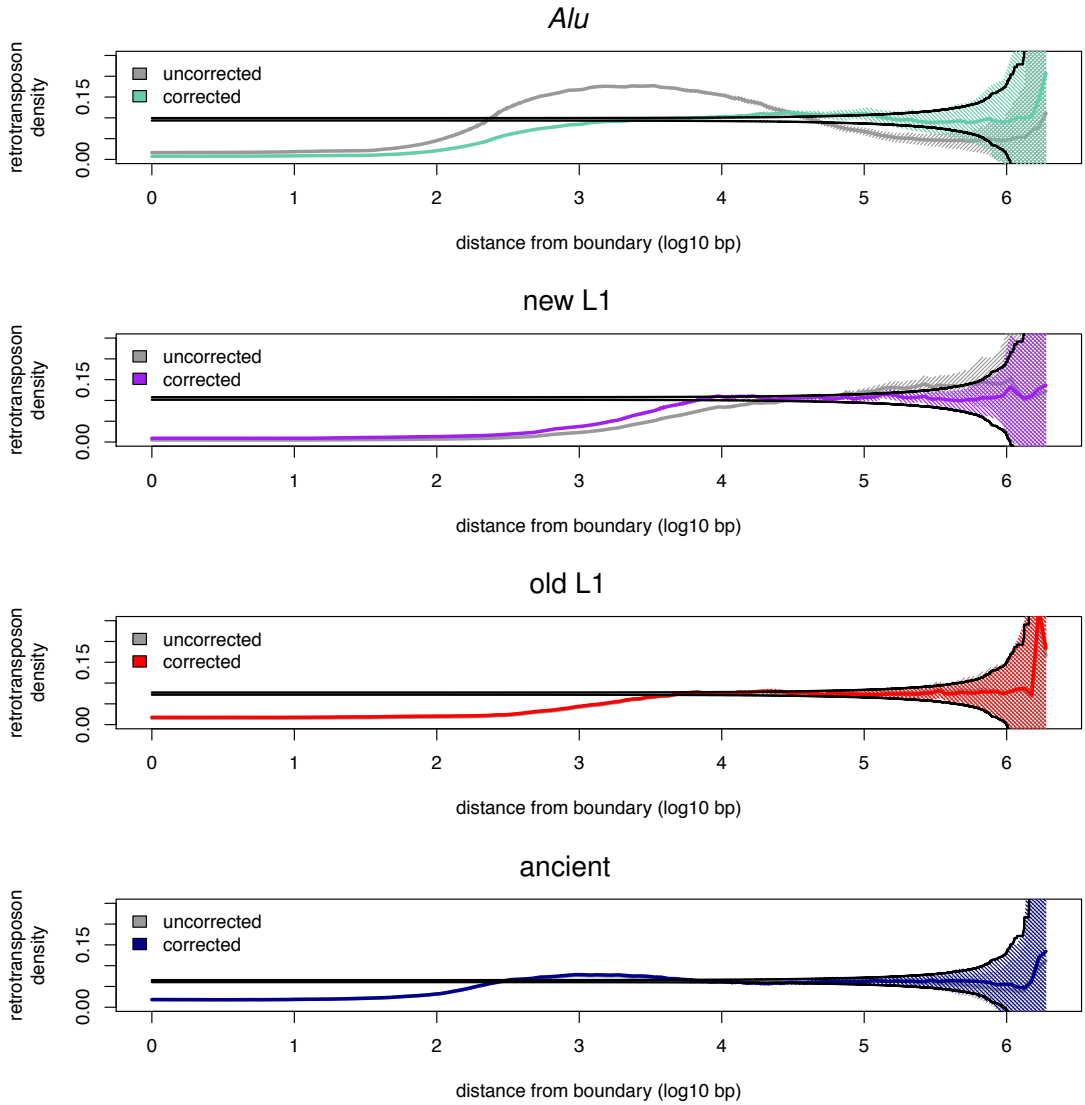


Figure S30: **Smoothed corrected retrotransposon density at the boundaries of genes.** Retrotransposon densities were corrected for bias levels corresponding to the interval sizes present at each position from the boundary. Black lines represent three standard deviations from expected densities. Retrotransposon densities were smoothed by using an expanding window from which variance was calculated. Retrotransposon density variance is represented by cross hatching at three standard deviations.

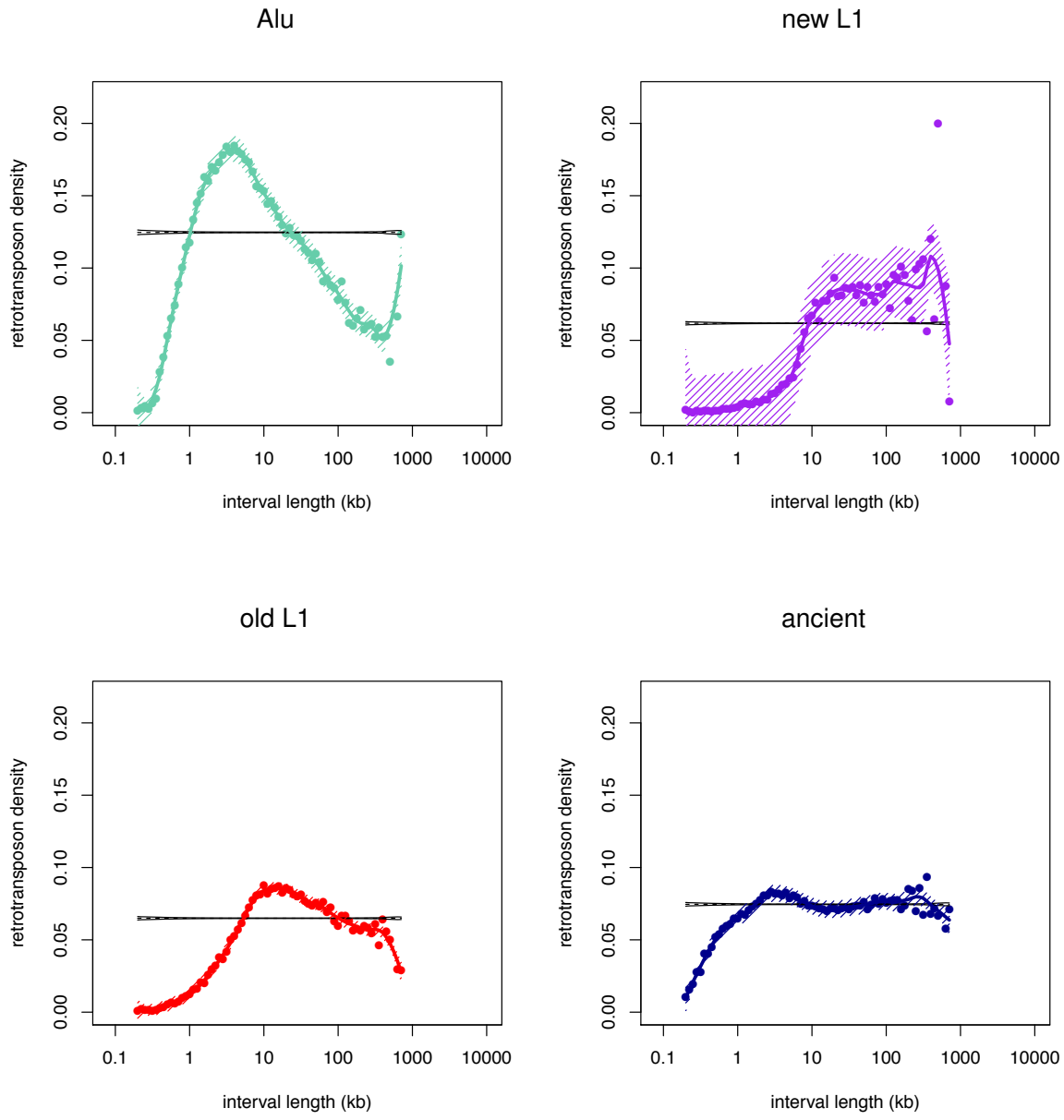


Figure S31: **Interval size bias of retrotransposon density for introns.** Retrotransposon density for different sized intervals is represented at each point. A line was fitted using LOESS and standard error estimates are represented by cross hatching. The expected Retrotransposon density at each interval size is shown by the dotted line. A confidence interval of 3 standard deviations from expected is represented by black solid lines.

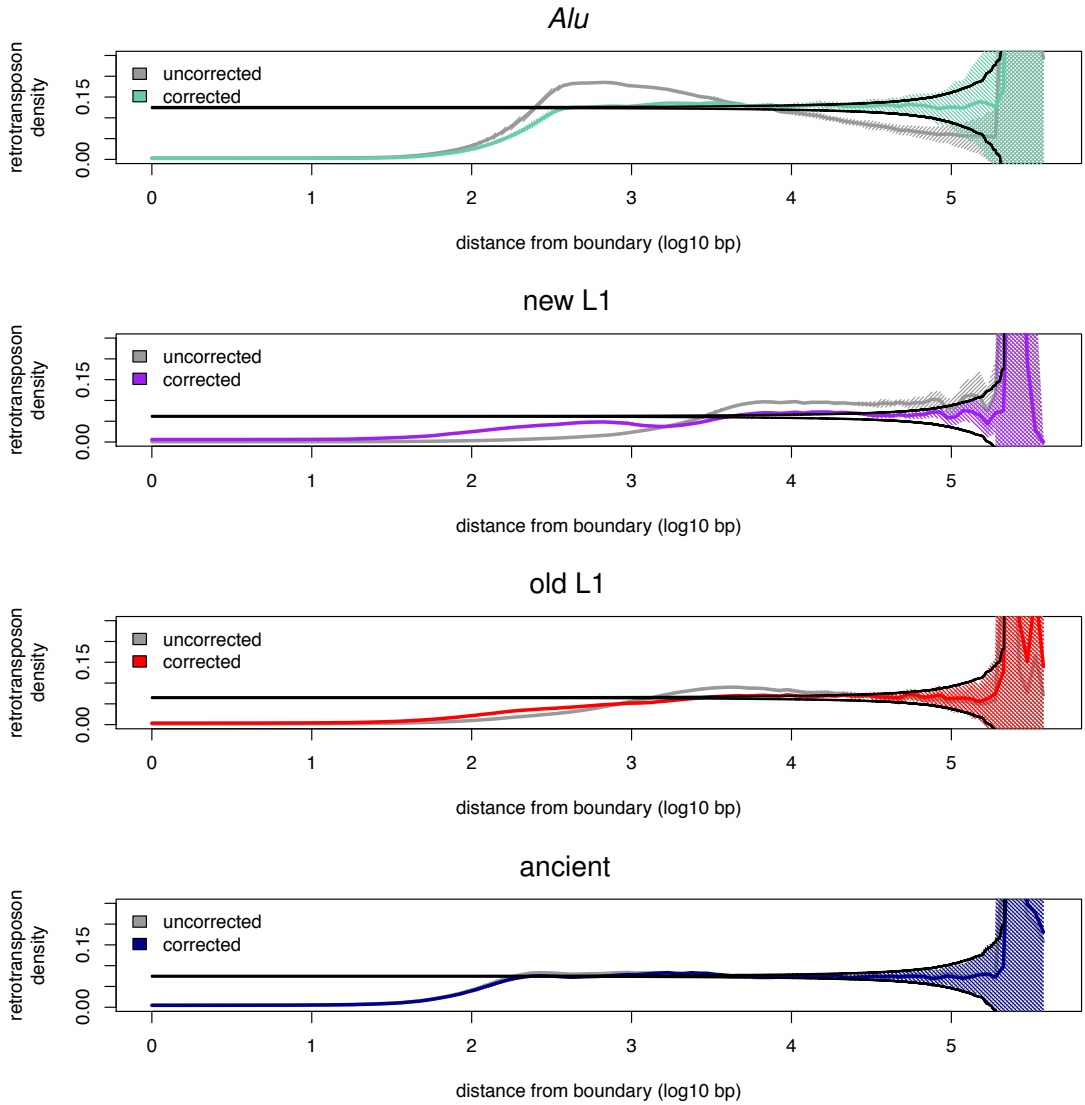


Figure S32: **Smoothed corrected retrotransposon density at the boundaries of exons.** Retrotransposon densities were corrected for bias levels corresponding to the interval sizes present at each position from the boundary. Black lines represent three standard deviations from expected densities. Retrotransposon densities were smoothed by using an expanding window from which variance was calculated. Retrotransposon density variance is represented by cross hatching at three standard deviations.

Supplementary Tables

Species	Assembly	source	Retrotransposon annotation url	AXT alignment url
human	hg19	UCSC	hg19/database/rmsk.txt.gz	
chimpanzee	panTro4	UCSC	panTro4/database/rmsk.txt.gz	hg19/vsPanTro4/axtNet/
rhesus macaque	rheMac3	UCSC	rheMac3/database/rmsk.txt.gz	hg19/vsRheMac3/axtNet/
mouse	mm9	UCSC	mm9/database/rmsk.txt.gz	hg19/vsMm9/axtNet/
rabbit	oryCun2	RepeatMasker	oryCun.html	hg19/vsOryCun2/axtNet/
dog	canFam3	UCSC	canFam3/database/rmsk.txt.gz	hg19/vsCanFam3/axtNet/
horse	equCab2	RepeatMasker	equCab.html	hg19/vsEquCab2/axtNet/
pig	susScr2	RepeatMasker	susScr.html	hg19/vsSusScr2/axtNet/

Table S1: **Datasets we extracted retrotransposon coordinates and genome alignments from.** For retrotransposon annotations, tables sourced from UCSC can be found by adding the extensions to the following url: `http://hgdownload.soe.ucsc.edu/goldenPath/`. Tables sourced form the RepeatMasker website can be found by adding the extensions to `http://www.repeatmasker.org/genomes/` and clicking on the appropriate `.fa.out.gz` file. For AXT alignments, the data can be found by adding the extension to `http://hgdownload.soe.ucsc.edu/goldenPath/`.

cell line	replication domain files	replication timing url extension
BG02ES	GSE53984_GSM923453_Bg02es_Rep1_segments.bed.gz	wgEncodeUwRepliSeqBg02esWaveSignalRep1.txt.gz
BJ	GSE53984_GSM923444_Bj_Rep1_segments.bed.gz	wgEncodeUwRepliSeqBjWaveSignalRep1.txt.gz
BJ	GSE53984_GSM923444_Bj_Rep2_segments.bed.gz	wgEncodeUwRepliSeqBjWaveSignalRep2.txt.gz
GM06990	GSE53984_GSM923443_Gm06990_Rep1_segments.bed.gz	wgEncodeUwRepliSeqGm06990WaveSignalRep1.txt.gz
GM12801	GSE53984_GSM923440_Gm12801_Rep1_segments.bed.gz	wgEncodeUwRepliSeqGm12801WaveSignalRep1.txt.gz
GM12812	GSE53984_GSM923439_Gm12812_Rep1_segments.bed.gz	wgEncodeUwRepliSeqGm12812WaveSignalRep1.txt.gz
GM12813	GSE53984_GSM923450_Gm12813_Rep1_segments.bed.gz	wgEncodeUwRepliSeqGm12813WaveSignalRep1.txt.gz
GM12878	GSE53984_GSM923451_Gm12878_Rep1_segments.bed.gz	wgEncodeUwRepliSeqGm12878WaveSignalRep1.txt.gz
HeLa-S3	GSE53984_GSM923449_Helas3_Rep1_segments.bed.gz	wgEncodeUwRepliSeqHelas3WaveSignalRep1.txt.gz
HepG2	GSE53984_GSM923446_Hepg2_Rep1_segments.bed.gz	wgEncodeUwRepliSeqHepg2WaveSignalRep1.txt.gz
HUVEC	GSE53984_GSM923452_Huvec_Rep1_segments.bed.gz	wgEncodeUwRepliSeqHuvecWaveSignalRep1.txt.gz
IMR90	GSE53984_GSM923447_Imr90_Rep1_segments.bed.gz	wgEncodeUwRepliSeqImr90WaveSignalRep1.txt.gz
K562	GSE53984_GSM923448_K562_Rep1_segments.bed.gz	wgEncodeUwRepliSeqK562WaveSignalRep1.txt.gz
MCF-7	GSE53984_GSM923442_Mcf7_Rep1_segments.bed.gz	wgEncodeUwRepliSeqMcf7WaveSignalRep1.txt.gz
NHEK	GSE53984_GSM923445_Nhek_Rep1_segments.bed.gz	wgEncodeUwRepliSeqNhekWaveSignalRep1.txt.gz
SK-N-SH	GSE53984_GSM923441_Sknsh_Rep1_segments.bed.gz	wgEncodeUwRepliSeqSknshWaveSignalRep1.txt.gz

Table S2: **Datasets we extracted human repli-Seq replication timing profiles from.** Replication domains can be found via the accession GSE53984 or the url: `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53984`. Replication timing profiles can be found on the UCSC genome browser by adding the extensions to the following url: `http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/`.

cell-line	replication timing url extension
Ch12F	wgEncodeFsuRepliChipCh12FWaveSignalRep1.bigWig
Ch12F	wgEncodeFsuRepliChipCh12FWaveSignalRep2.bigWig
Episc5M	wgEncodeFsuRepliChipEpisc5MWaveSignalRep1.bigWig
Episc5M	wgEncodeFsuRepliChipEpisc5MWaveSignalRep2.bigWig
Episc7F	wgEncodeFsuRepliChipEpisc7FWaveSignalRep1.bigWig
Episc7F	wgEncodeFsuRepliChipEpisc7FWaveSignalRep2.bigWig
Esd3MDiffe3d	wgEncodeFsuRepliChipEsd3MDiffe3dWaveSignalRep1.bigWig
Esd3MDiffe3d	wgEncodeFsuRepliChipEsd3MDiffe3dWaveSignalRep2.bigWig
Esd3MDiffe6d	wgEncodeFsuRepliChipEsd3MDiffe6dWaveSignalRep1.bigWig
Esd3MDiffe6d	wgEncodeFsuRepliChipEsd3MDiffe6dWaveSignalRep2.bigWig
Esd3MDiffe9d	wgEncodeFsuRepliChipEsd3MDiffe9dWaveSignalRep1.bigWig
Esd3MDiffe9d	wgEncodeFsuRepliChipEsd3MDiffe9dWaveSignalRep2.bigWig
Esd3MDiffg3d	wgEncodeFsuRepliChipEsd3MDiffg3dWaveSignalRep1.bigWig
Esd3MDiffg3d	wgEncodeFsuRepliChipEsd3MDiffg3dWaveSignalRep2.bigWig
Esd3M	wgEncodeFsuRepliChipEsd3MWaveSignalRep1.bigWig
Esd3M	wgEncodeFsuRepliChipEsd3MWaveSignalRep2.bigWig
Esem5sUDiffhsoxm	wgEncodeFsuRepliChipEsem5sUDiffhsoxmWaveSignalRep1.bigWig
Esem5sUDiffhsoxm	wgEncodeFsuRepliChipEsem5sUDiffhsoxmWaveSignalRep2.bigWig
Esem5sUDiffhsoxp	wgEncodeFsuRepliChipEsem5sUDiffhsoxpWaveSignalRep1.bigWig
Esem5sUDiffhsoxp	wgEncodeFsuRepliChipEsem5sUDiffhsoxpWaveSignalRep2.bigWig
Estt2MDifff9d	wgEncodeFsuRepliChipEstt2MDifff9dWaveSignalRep1.bigWig
Estt2MDifff9d	wgEncodeFsuRepliChipEstt2MDifff9dWaveSignalRep2.bigWig
Estt2M	wgEncodeFsuRepliChipEstt2MWaveSignalRep1.bigWig
Estt2M	wgEncodeFsuRepliChipEstt2MWaveSignalRep2.bigWig
J185aU	wgEncodeFsuRepliChipJ185aUWaveSignalRep1.bigWig
J185aU	wgEncodeFsuRepliChipJ185aUWaveSignalRep2.bigWig
L1210F	wgEncodeFsuRepliChipL1210FWaveSignalRep1.bigWig
L1210F	wgEncodeFsuRepliChipL1210FWaveSignalRep2.bigWig
MelM	wgEncodeFsuRepliChipMelMWaveSignalRep1.bigWig
MelM	wgEncodeFsuRepliChipMelMWaveSignalRep2.bigWig

Table S3: **Datasets we extracted mouse repli-CHiP replication timing profiles from.** Replication timing profiles can be found on the UCSC genome browser by adding the extensions to the following url: `http://hgdownload.cse.ucsc.edu/gbdb/mm9/bbi/`.

cell line	url extension
Gliobla	wgEncodeOpenChromSynthGlioblaPk.txt.gz
GM12878	wgEncodeOpenChromSynthGm12878Pk.txt.gz
GM12891	wgEncodeOpenChromSynthGm12891Pk.txt.gz
GM12892	wgEncodeOpenChromSynthGm12892Pk.txt.gz
GM18507	wgEncodeOpenChromSynthGm18507Pk.txt.gz
GM19239	wgEncodeOpenChromSynthGm19239Pk.txt.gz
H1-hESC	wgEncodeOpenChromSynthH1hescPk.txt.gz
HeLa-S3	wgEncodeOpenChromSynthHelas3Pk.txt.gz
HepG2	wgEncodeOpenChromSynthHepg2Pk.txt.gz
HTR8svn	wgEncodeOpenChromSynthHtr8Pk.txt.gz
HUVEC	wgEncodeOpenChromSynthHuvecPk.txt.gz
K562	wgEncodeOpenChromSynthK562Pk.txt.gz
Medullo	wgEncodeOpenChromSynthMedulloPk.txt.gz
NHEK	wgEncodeOpenChromSynthNhekPk.txt.gz
PanIslets	wgEncodeOpenChromSynthPanisletsPk.txt.gz

Table S4: **Datasets we extracted DNase1 hypersensitive sites from.** Tables can be found on the UCSC genome browser by adding the extensions to the following url: `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/`.

Intergenic		**Intron**	
filtering operation	intervals	filtering operation	intervals
extract intervals between genes	21,392	extract intervals between exons	505,857
remove intervals adjacent chromosome edges	21,359	remove alternatively spliced intervals	170,594
remove intervals with strand conflicts	20,148	remove intervals overlapping other genes	164,448

Table S5: **Filtering of intergenic and intron intervals.** All operations were performed on RefSeq gene annotations obtained from the UCSC genome browser.

Cell	Lineage	Tissue	Sex	DNase1	Repli.Seq
BG02ES	inner cell mass	embryonic stem cell	M		✓
BJ		skin	M		✓
GM06990	mesoderm	blood	F		✓
GM12801	mesoderm	blood	M		✓
GM12812	mesoderm	blood	M		✓
GM12813	mesoderm	blood	F		✓
GM12878	mesoderm	blood	F		✓
GM12891	mesoderm	blood	M	✓	
GM12892	mesoderm	blood	F	✓	
GM18507	mesoderm	blood	M	✓	
GM19239	mesoderm	blood	M	✓	
H1-hESC	inner cell mass	embryonic stem cell	M	✓	
HeLa-S3	ectoderm	cervix	F	✓	✓
HepG2	endoderm	liver	M	✓	✓
HTR8svn	ectoderm	blastula	F	✓	
HUVEC	mesoderm	blood vessel	U	✓	✓
IMR90	endoderm	lung	F		✓
K562	mesoderm	blood	F	✓	✓
MCF-7	ectoderm	breast	F		✓
Medullo	ectoderm	brain	U	✓	
NHEK	ectoderm	skin	U	✓	✓
PanIslets	endoderm	pancreas	M	✓	
SK-N-SH	ectoderm	brain	F		✓

Table S6: **Cell-line information for human ENCODE data.** Information was obtained from `https://genome.ucsc.edu/ENCODE/cellTypes.html`.

Source	Description	Catagory	Tissue	Sex
CH12	B-cell lymphoma (GM12878 analog)	cellLine	blood	F
EpiSC-5	epidermal stem cell	primaryCells		M
EpiSC-7	epidermal stem cell	primaryCells		M
ES-D3	ES-cells isolated from 129S2/SvPas	primaryCells		M
ES-EM5Sox17huCD25	ES cell line that bears the gfp and human IL2R alpha (also known as CD25) marker genes in the goosecoid (Gsc) and Sox17 loci, derived from EB5	primaryCells		U
ES-TT2	ES-cells isolated from C57BL/6xCBA	primaryCells		M
J185a	Fetal myoblast Desmin+	cellLine		U
L1210	lymphoblast from 8 month female	cellLine	blood	F
MEL	Leukemia (K562 analog)	cellLine	blood	M

Table S7: **Cell-line information for mouse ENCODE data.** Information was obtained from `https://genome.ucsc.edu/ENCODE/cellTypesMouse.html`.

Human

	Sknsh	Nhek	Mcf7	K562	Imr90	Huvec	Hepg2	Helas3	Gm12878	Gm12813	Gm12812	Gm12801	Gm06990	Bj	Bg02es
Bg02es	0.68	0.73	0.7	0.73	0.68	0.73	0.76	0.69	0.71	0.7	0.7	0.72	0.72	0.69	1
Bj	0.72	0.87	0.75	0.75	0.94	0.86	0.79	0.8	0.79	0.78	0.78	0.79	0.79	1	
Gm06990	0.76	0.85	0.77	0.84	0.77	0.78	0.84	0.75	0.98	0.95	0.94	0.97	1		
Gm12801	0.75	0.84	0.77	0.85	0.76	0.79	0.85	0.75	0.97	0.97	0.98	1			
Gm12812	0.72	0.82	0.75	0.83	0.75	0.78	0.82	0.75	0.95	0.97	1				
Gm12813	0.75	0.84	0.77	0.83	0.77	0.78	0.83	0.75	0.97	1					
Gm12878	0.77	0.85	0.77	0.85	0.77	0.78	0.84	0.75	1						
Helas3	0.66	0.81	0.77	0.72	0.78	0.81	0.77	1							
Hepg2	0.76	0.84	0.81	0.85	0.78	0.8	1								
Huvec	0.71	0.84	0.76	0.78	0.85	1									
Imr90	0.74	0.86	0.74	0.74	1										
K562	0.74	0.81	0.78	1											
Mcf7	0.72	0.81	1												
Nhek	0.77	1													
Sknsh	1														

Mouse

	MelM	L1210F	J185aU	Estt2M	Estt2MDifff9d	Esem5sUDiffhsoxp	Esem5sUDiffhsoxm	Esd3M	Esd3MDiffg3d	Esd3MDiffe9d	Esd3MDiffe6d	Esd3MDiffe3d	Episc7F	Episc5M	Ch12F
Ch12F	0.8	0.73	0.65	0.66	0.66	0.64	0.64	0.68	0.67	0.65	0.65	0.62	0.63	0.68	1
Episc5M	0.69	0.83	0.7	0.89	0.92	0.78	0.78	0.88	0.78	0.86	0.93	0.9	0.95	1	
Episc7F	0.65	0.83	0.64	0.89	0.93	0.71	0.72	0.86	0.71	0.85	0.95	0.94	1		
Esd3MDiffe3d	0.63	0.79	0.61	0.93	0.92	0.7	0.69	0.89	0.74	0.85	0.94	1			
Esd3MDiffe6d	0.64	0.84	0.68	0.91	0.97	0.75	0.75	0.87	0.75	0.94	1				
Esd3MDiffe9d	0.59	0.79	0.69	0.85	0.93	0.7	0.7	0.84	0.74	1					
Esd3MDiffg3d	0.64	0.69	0.82	0.78	0.74	0.89	0.88	0.81	1						
Esd3M	0.65	0.8	0.67	0.96	0.86	0.73	0.72	1							
Esem5sUDiffhsoxm	0.66	0.68	0.81	0.72	0.74	0.95	1								
Esem5sUDiffhsoxp	0.65	0.68	0.8	0.72	0.73	1									
Estt2MDifff9d	0.65	0.82	0.68	0.91	1										
Estt2M	0.65	0.81	0.66	1											
J185aU	0.61	0.67	1												
L1210F	0.69	1													
MelM	1														

Table S8: **Pairwise genomic correlations for cell-line replication timing.** Pairwise correlation analysis was carried out on mean replication timing per 1 Mb genome segments.

Appendix B

Supplementary for Chapter 3

Supplementary figures

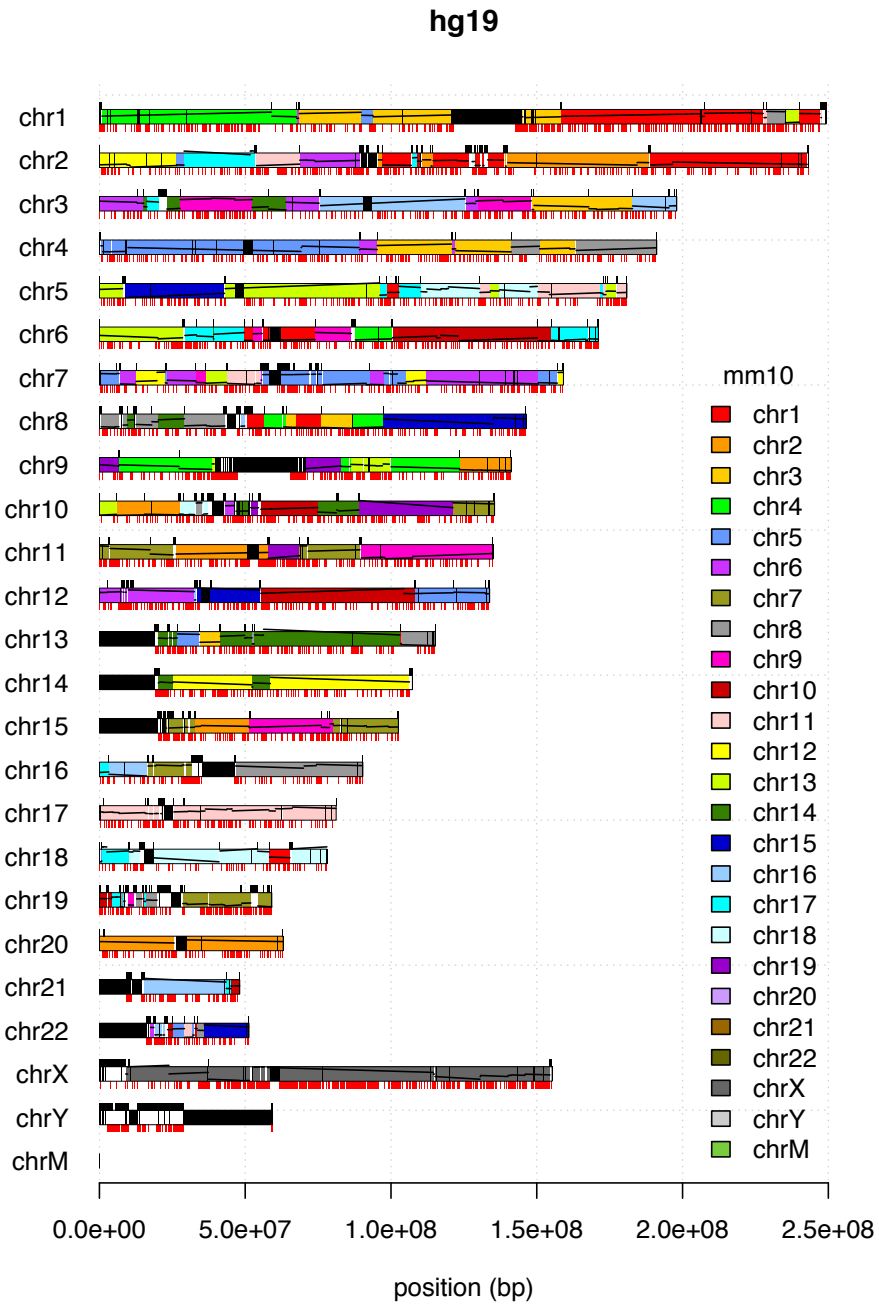


Figure S1: Genomic regions filtered from hg19. Gaps outside of nets $\geq$ 10 kb are shown in black above each chromosome. non-RBH regions $\geq$ 10 kb are shown in red below each chromosome. Assembly gaps are plotted in black within chromosomes. Syntenic blocks are coloured according to which chromosome they belong to in mm10. The trace running through each syntenic block represents its mm10 chromosomal position and orientation, running top to bottom (5′ to 3′).

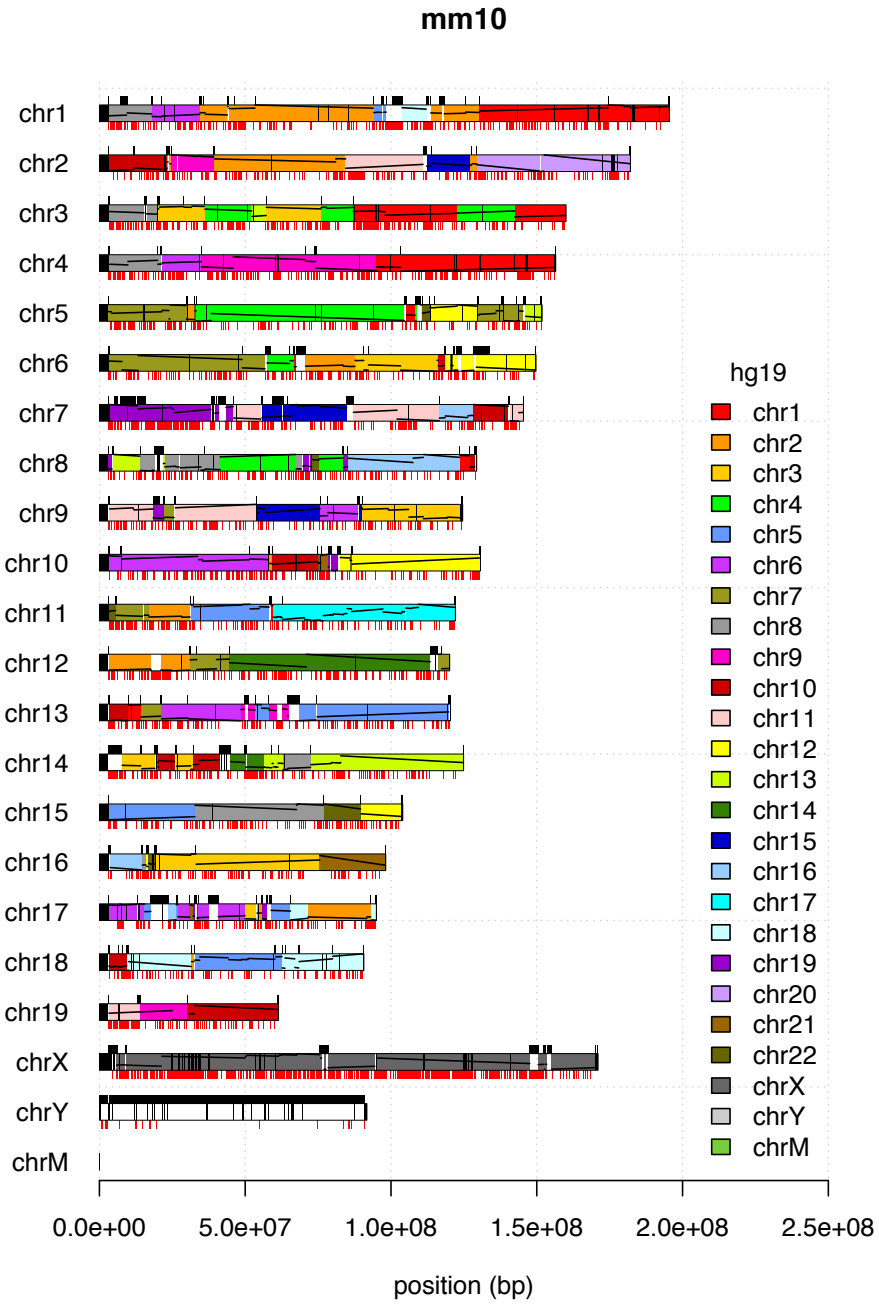


Figure S2: Genomic regions filtered from mm10. Gaps outside of nets $\geq$ 10 kb are shown in black above each chromosome. non-RBH regions $\geq$ 10 kb are shown in red below each chromosome. Assembly gaps are plotted in black within chromosomes. Syntenic blocks are coloured according to which chromosome they belong to in hg19. The trace running through each syntenic block represents its hg19 chromosomal position and orientation, running top to bottom ($5'$ to $3'$).

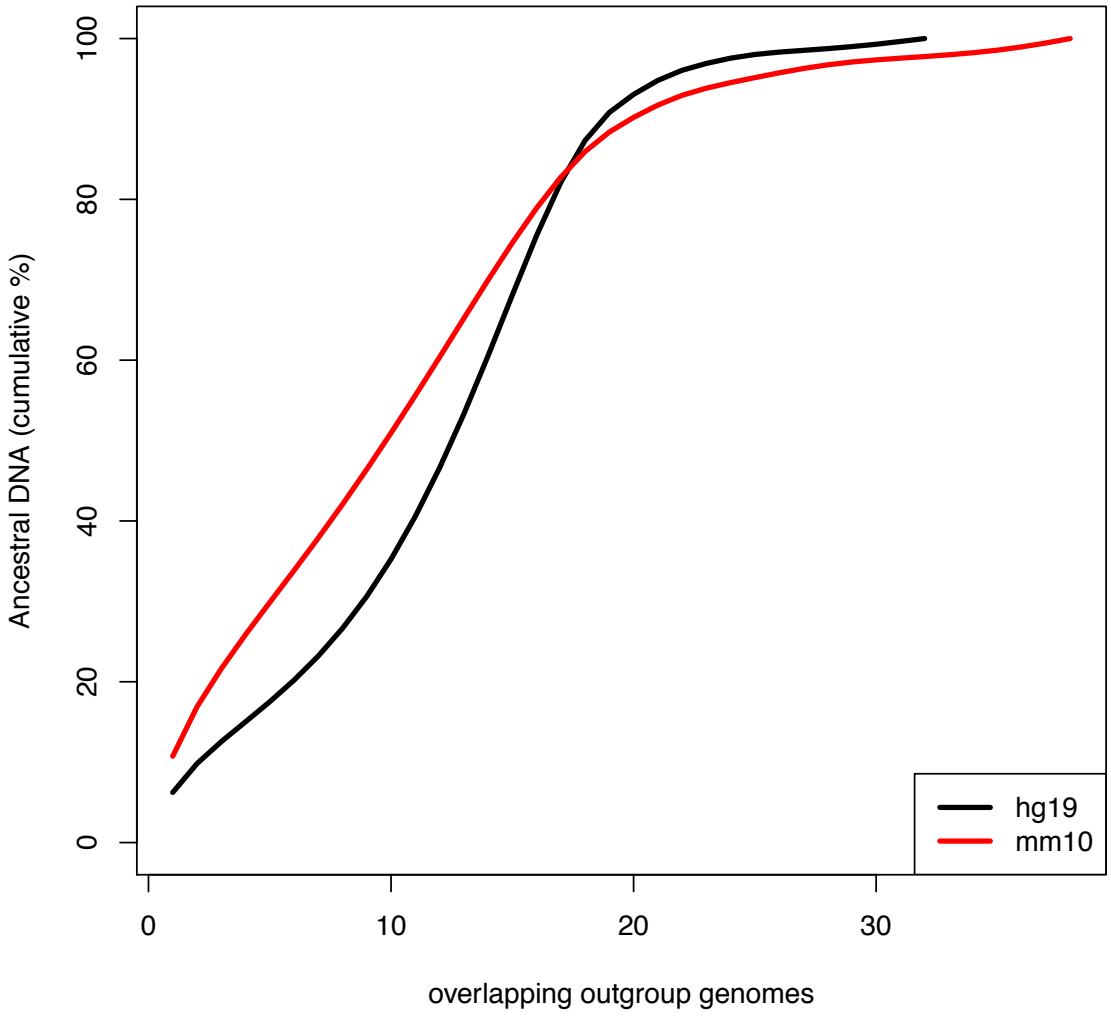


Figure S3: Coverage depth of fills extracted from outgroup species. Coverage depth is measured by number of overlapping outgroup species. Ancestral DNA % is the proportion of total bp in hg19 and mm10 that overlap at least one fill extracted from an outgroup species.

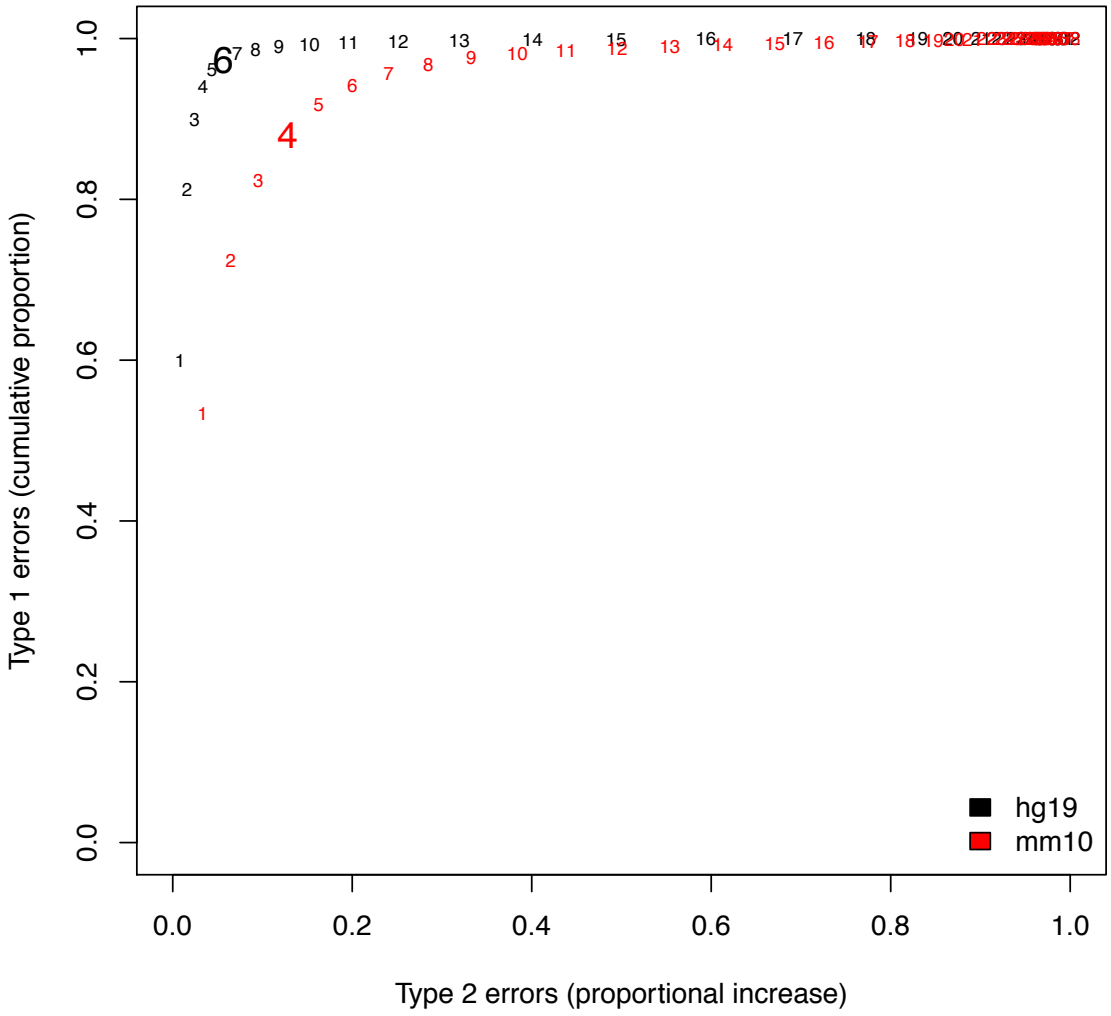


Figure S4: Error profile and coverage depth for identifying ancestral elements. Minimum coverage depth threshold for identifying ancestral elements is plotted against total proportion of identified type 1 errors and the proportional increase in type 2 error rate. Type 1 errors are identified as known recent transposons that overlap fills extracted from outgroup species. Type 2 errors are identified as fills between hg19 and mm10 that do not overlap fills extracted from outgroup species. Type 2 error increase is the reduction in the overlap between outgroup and ingroup (hg19 and mm10) fills as minnimum coverage depth threshold increases. For hg19 and mm10 we chose a minimum coverage depth of 6 and 4 respectively.

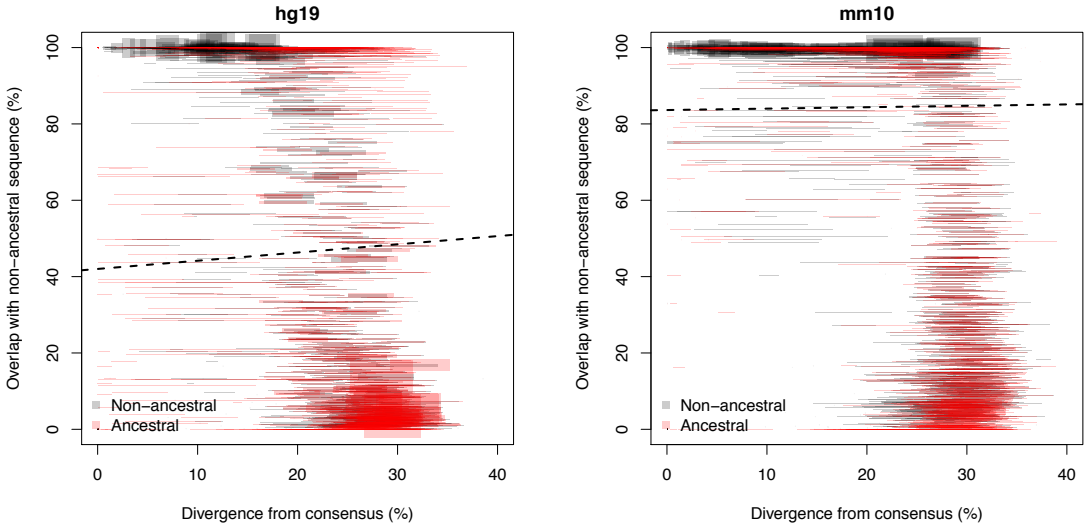


Figure S5: Transposon family classification with linear discriminant analysis. Each rectangle represents the members of a transposon family under our prior recent and ancestral classification. For example, a rectangle coloured black represents the members of a particular transposon family that do not overlap ancestral elements. Rectangle width is the interquartile range of percent divergence from consensus and rectangle height is proportional to total genome coverage. The dotted line is the classification boundary determined by linear discriminant analysis. Rectangles above the line are transposon families classified as recent and rectangles below the line are transposon families classified as ancestral.

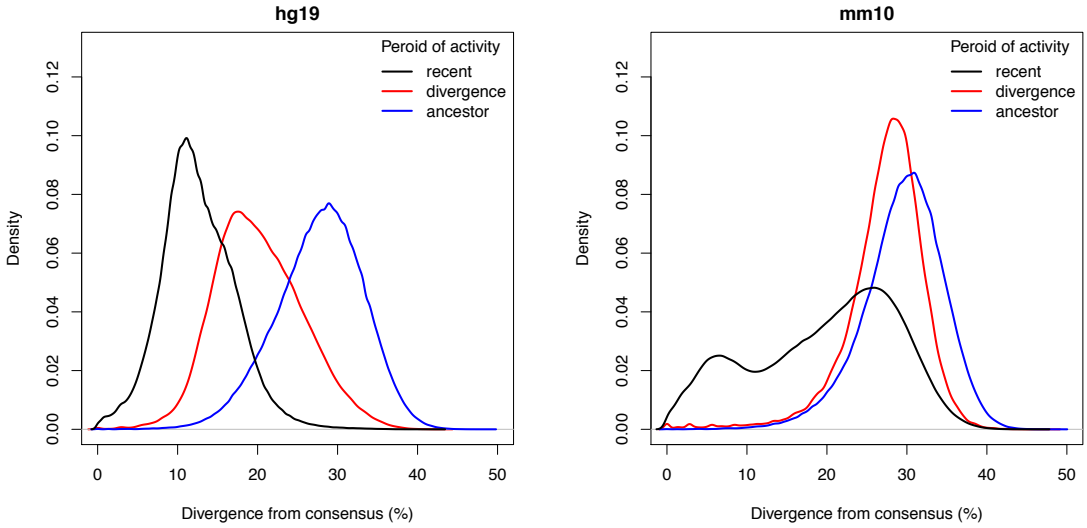


Figure S6: Transposon family period of activity and percent divergence from consensus. Transposons identified as recently active were classified as recent by our classifier and belong to families not shared between human and mouse. Transposons identified as active during divergence were classified as recent by our classifier and belong to families shared between human and mouse. Transposons identified as active within the ancestor were classified as ancestral by our classifier and belong to families shared between human and mouse. Transposons classified as ancestral by our classifier that belong to families not shared between human and mouse are not shown.

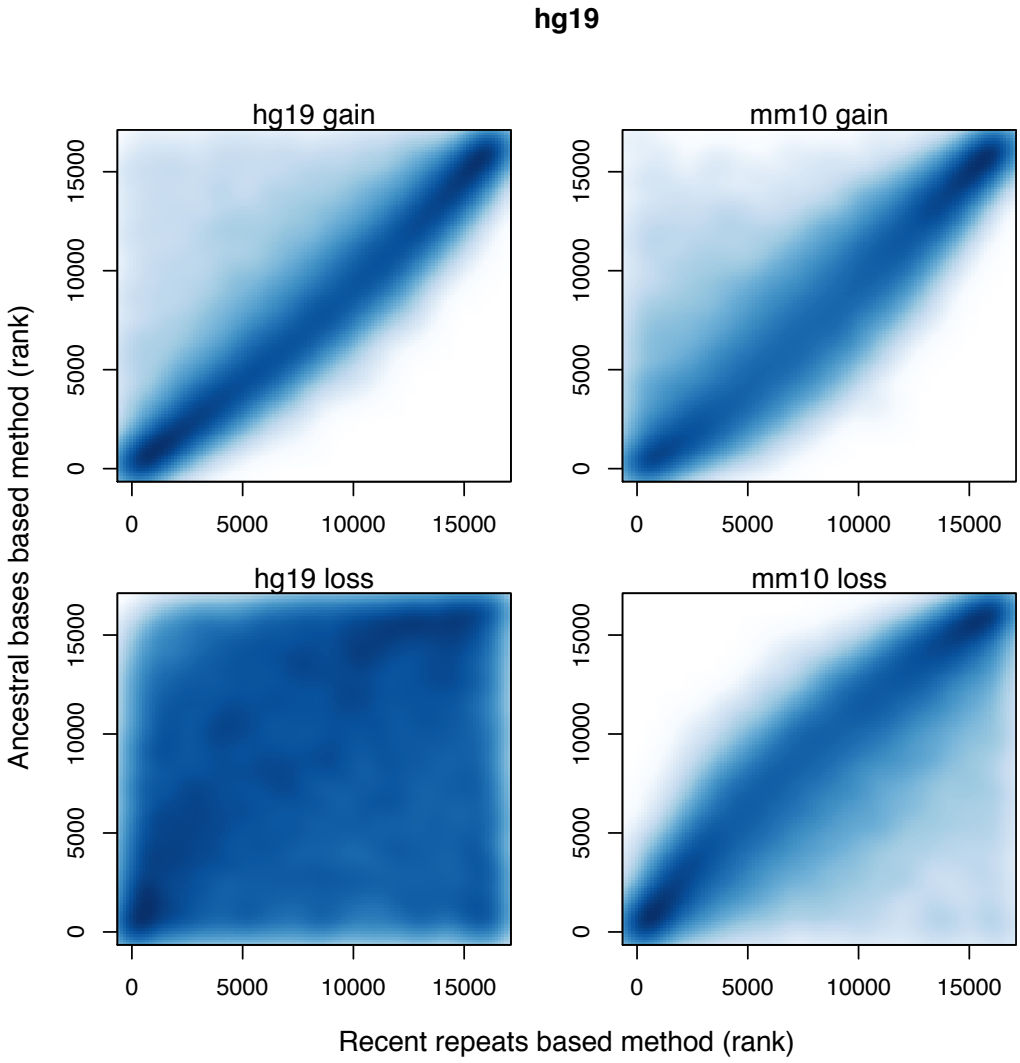


Figure S7: Rank comparison of gap annotation methods per 200 kb bin in hg19 genomic background.

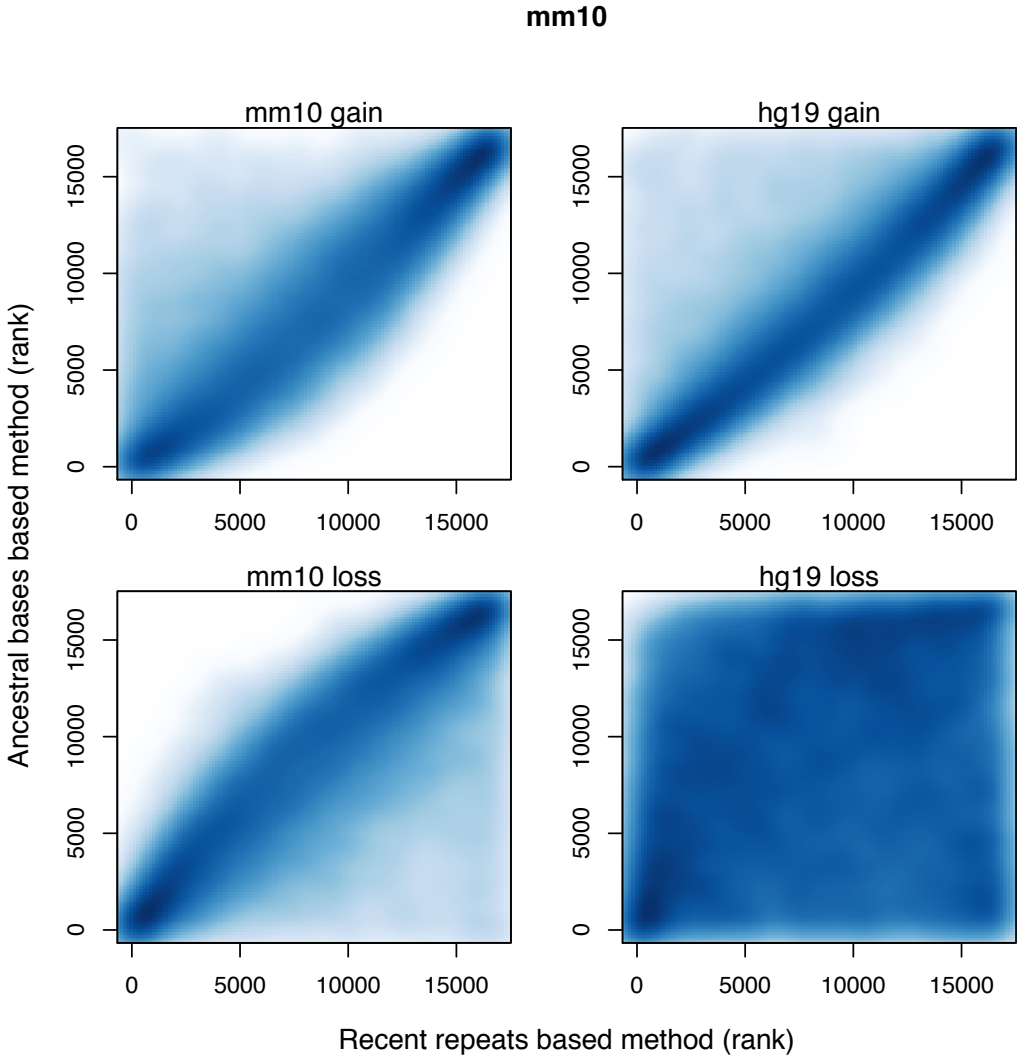
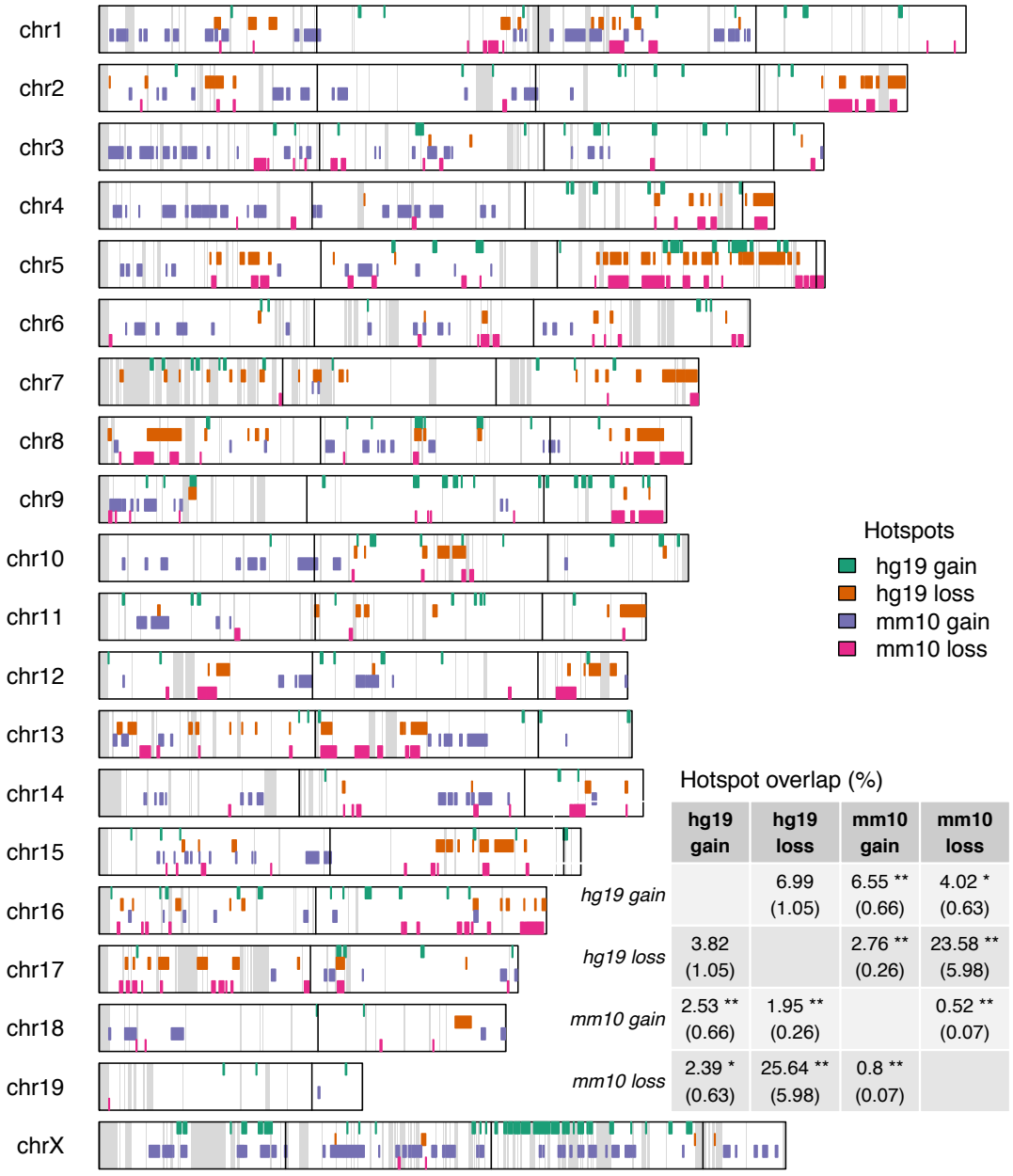


Figure S8: Rank comparison of gap annotation methods per 200 kb bin in mm10 genomic background.



Hotspot overlap (%)

	hg19 gain	hg19 loss	mm10 gain	mm10 loss
hg19 gain		6.99 (1.05)	6.55 ** (0.66)	4.02 * (0.63)
hg19 loss	3.82 (1.05)		2.76 ** (0.26)	23.58 ** (5.98)
mm10 gain	2.53 ** (0.66)	1.95 ** (0.26)		0.52 ** (0.07)
mm10 loss	2.39 * (0.63)	25.64 ** (5.98)	0.8 ** (0.07)	

Figure S9: Genomic distribution of gain and loss hotspots for hg19 and mm10 plotted against mm10 synthetic genome. Grey regions indicate bins with $\leq$ 150 kb of RBH nets and black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss hotspots. The percentages were calculated using the hotspots labelled in each row as the denominator. '*' and '**' represent p-values below .05 and .01 respectively based on the Fisher statistic.

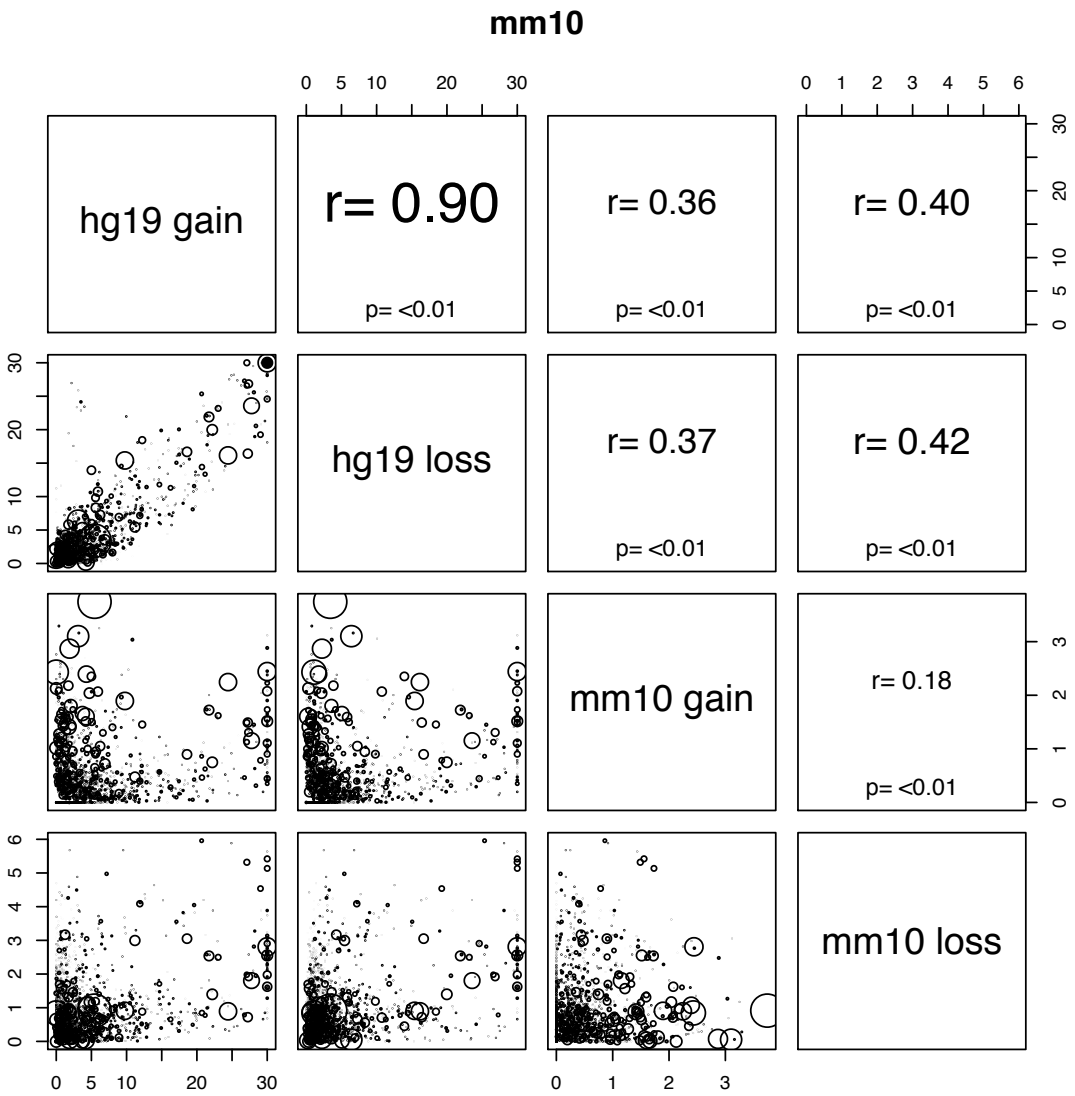


Figure S10: Over representation of biological process GO terms in gain and loss hotspots in mm10. The axes are marked according to -log10 P-values. The size of points represents the total umber of annotations for each GO term.

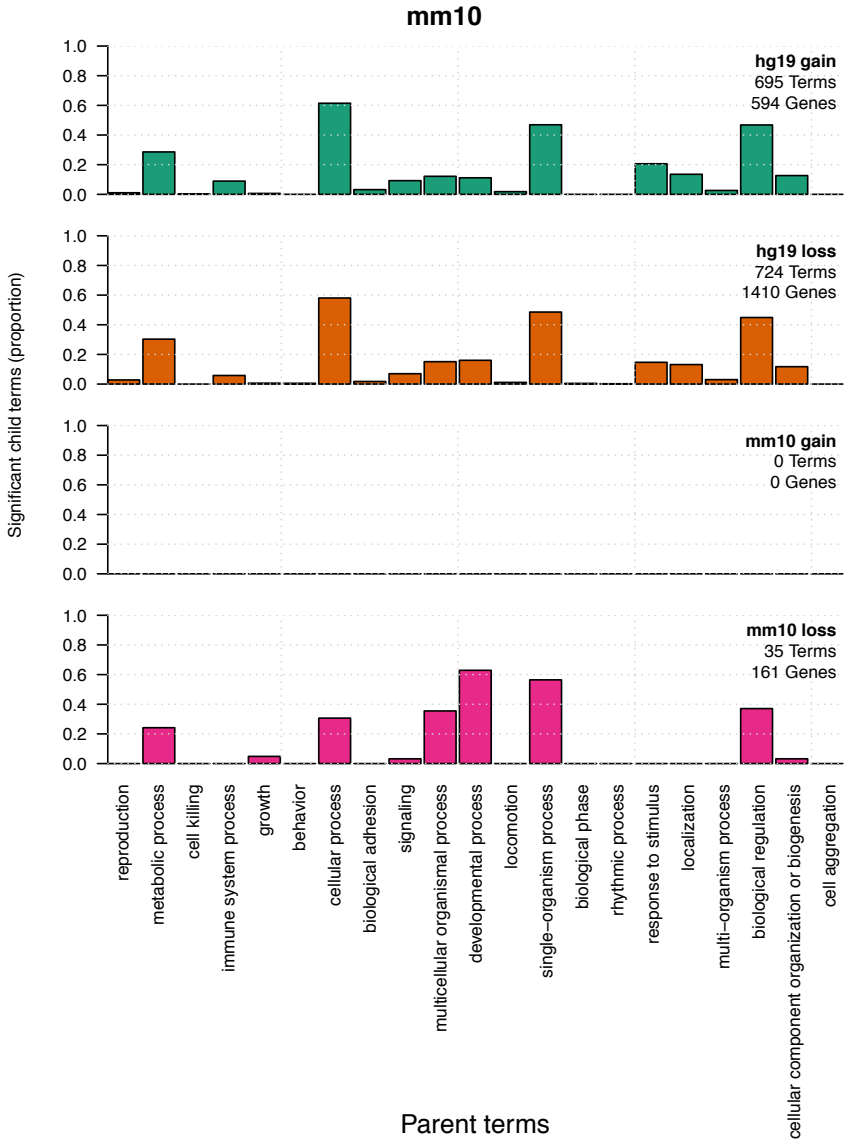


Figure S11: Significant biological process GO terms in mm10 background. Parent terms were the top level biological process GO terms while child terms were those beneath each parent term. Child terms were identified as significant at a FDR < 0.05 based on a Fisher test using the 'classic' algorithm. The Y axis represents the proportion of child GO terms that belong to each parant GO term. Proportions don't add up to 1 because some child GO terms are shared between parent GO terms. We also show the number of non-redundant GO terms and genes annotated with significant GO terms for each gap annotation.

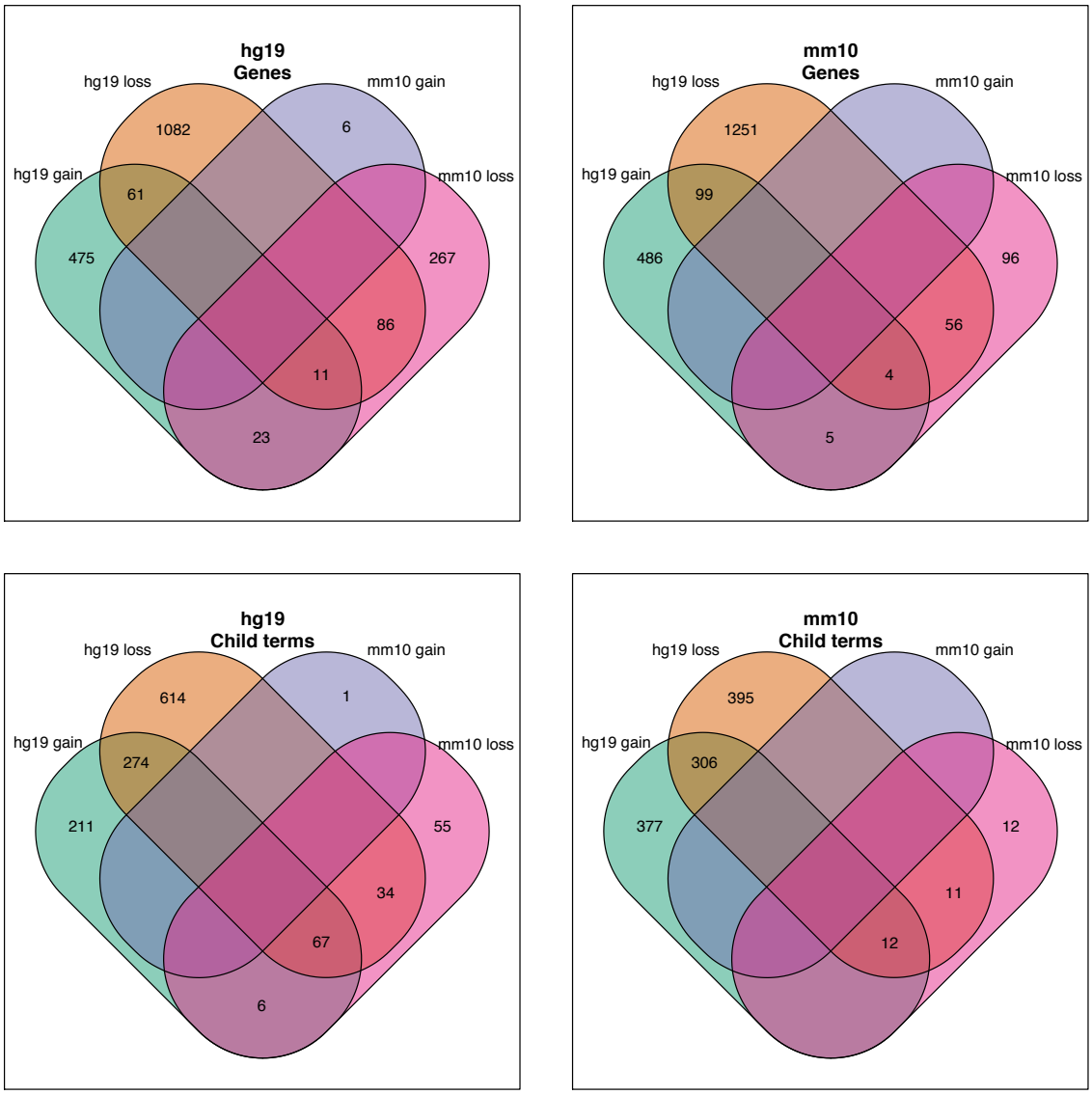


Figure S12: Comparison of significant biological process GO terms and annotated genes. GO terms were identified as significant at a FDR < 0.05 based on a Fisher test using the 'classic' algorithm. Annotated genes are genes that have been annotated with at least one of the significant GO terms. GO term lists and gene lists in each set are non-redundant.

Supplementary tables

hg19 transposon classification				
		LDA classification		
		Recent	Ancestral	**Total**
Family name	Lineage-specific	435.5 (480)	0.03 (6)	435.5 (486)
	Shared	136.1 (176)	345.4 (660)	481.5 (836)
	Total	571.6 (656)	345.4 (666)	917.0 (1322)
mm10 transposon classification				
		LDA classification		
		Recent	Ancestral	**Total**
Family name	Lineage-specific	636.7 (512)	0.4 (5)	637.1 (517)
	Shared	27.6 (177)	71.3 (659)	98.9 (836)
	Total	664.3 (689)	71.7 (664)	736.0 (1353)

Table S1: hg19 and mm10 classification of transposon families. Transposon classification compares our LDA classifier against shared and lineage-specific transposon family names. Presented is the total Mb transposon coverage with number of families in brackets

	UCSC ID	hg19	mm10
1	ailMel1	✓	✓
2	allMis1	✓	
3	anoCar2	✓	✓
4	bosTau7	✓	
5	bosTau8		✓
6	canFam3	✓	✓
7	cerSim1	✓	✓
8	choHof1	✓	✓
9	chrPic1		✓
10	danRer10	✓	✓
11	dasNov2	✓	
12	dasNov3		✓
13	echTel2	✓	✓
14	equCab2	✓	✓
15	eriEur1		✓
16	eriEur2	✓	
17	felCat5	✓	
18	fr2	✓	
19	fr3		✓
20	gadMor1		✓
21	galGal3	✓	
22	galGal5		✓
23	gasAcu1	✓	✓
24	latCha1		✓
25	loxAfr3	✓	✓
26	macEug1	✓	
27	macEug2		✓
28	melGal1		✓
29	monDom5	✓	✓
30	myoLuc1	✓	
31	myoLuc2		✓
32	oreNil2		✓
33	ornAna1		✓
34	oryLat2	✓	✓
35	oviAri1		✓
36	oviAri3	✓	
37	petMar2	✓	✓
38	proCap1	✓	✓
39	pteVam1	✓	✓
40	sarHar1	✓	✓
41	sorAra2	✓	✓
42	susScr2	✓	
43	susScr3		✓
44	taeGut1		✓
45	tetNig2	✓	✓
46	triMan1		✓
47	turTru1	✓	
48	turTru2		✓
49	vicPac2	✓	✓
50	xenTro3	✓	✓

Table S2: List of outgroup genomes used to identify ancestral elements in hg19 and mm10

Algorithm	GO.ID	Term	Significant	Expected	p-value
classic	GO:0008150	biological_process	636	6.82	$< 1 \times 10^{-30}$
	GO:0009987	cellular process	581	6.15	$< 1 \times 10^{-30}$
	GO:0008152	metabolic process	438	4.48	$< 1 \times 10^{-30}$
	GO:0071704	organic substance metabolic process	428	4.32	$< 1 \times 10^{-30}$
	GO:0044763	single-organism cellular process	464	4.94	$< 1 \times 10^{-30}$
	GO:0044238	primary metabolic process	412	4.13	$< 1 \times 10^{-30}$
	GO:0044237	cellular metabolic process	412	4.14	$< 1 \times 10^{-30}$
	GO:0044699	single-organism process	494	5.45	$< 1 \times 10^{-30}$
	GO:0050794	regulation of cellular process	402	4.10	$< 1 \times 10^{-30}$
	GO:0065007	biological regulation	431	4.60	$< 1 \times 10^{-30}$
elim	GO:0045944	positive regulation of transcription fro...	46	0.45	$< 1 \times 10^{-30}$
	GO:0000122	negative regulation of transcription fro...	37	0.32	$< 1 \times 10^{-30}$
	GO:0051301	cell division	33	0.25	$< 1 \times 10^{-30}$
	GO:0007165	signal transduction	230	2.33	$< 1 \times 10^{-30}$
	GO:0043547	positive regulation of GTPase activity	25	0.28	$< 1 \times 10^{-30}$
	GO:0015031	protein transport	84	0.79	$< 1 \times 10^{-30}$
	GO:0055114	oxidation-reduction process	46	0.41	$< 1 \times 10^{-30}$
	GO:0008150	biological_process	636	6.82	$< 1 \times 10^{-30}$
	GO:0006355	regulation of transcription, DNA-templat...	131	1.38	1.5×10^{-30}
	GO:0008283	cell proliferation	90	0.83	2.6×10^{-30}
weight	GO:0006468	protein phosphorylation	92	0.82	1.2×10^{-16}
	GO:0006357	regulation of transcription from RNA pol...	85	0.79	1.5×10^{-15}
	GO:1903047	mitotic cell cycle process	59	0.40	5.1×10^{-14}
	GO:0034645	cellular macromolecule biosynthetic proc...	208	1.97	4.4×10^{-12}
	GO:0007165	signal transduction	230	2.33	8.5×10^{-12}
	GO:0010628	positive regulation of gene expression	74	0.73	5.7×10^{-11}
	GO:0065003	macromolecular complex assembly	83	0.71	8.3×10^{-11}
	GO:1902589	single-organism organelle organization	76	0.74	3.6×10^{-9}
	GO:0008283	cell proliferation	90	0.83	5.3×10^{-9}
	GO:0018193	peptidyl-amino acid modification	70	0.52	7.1×10^{-9}
parent child	GO:0043618	regulation of transcription from RNA pol...	11	0.03	1.4×10^{-4}
	GO:0043620	regulation of DNA-templated transcriptio...	11	0.03	1.5×10^{-4}
	GO:0072331	signal transduction by p53 class mediato...	24	0.11	1.9×10^{-4}
	GO:1901796	regulation of signal transduction by p53...	17	0.07	5.5×10^{-4}
	GO:0015031	protein transport	84	0.79	7.9×10^{-4}
	GO:0048172	regulation of short-term neuronal synapt...	5	0.01	1.6×10^{-3}
	GO:0019395	fatty acid oxidation	10	0.05	2.6×10^{-3}
	GO:0061337	cardiac conduction	8	0.06	2.7×10^{-3}
	GO:1903047	mitotic cell cycle process	59	0.40	2.9×10^{-3}
	GO:0007049	cell cycle	92	0.75	2.9×10^{-3}

Table S3: Top 10 biological process GO terms for genes located in hg19 gain hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in methods)

Algorithm	GO.ID	Term	Significant	Expected	p-value
classic	GO:0008150	biological_process	1365	14.64	$< 1 \times 10^{-30}$
	GO:0009987	cellular process	1229	13.20	$< 1 \times 10^{-30}$
	GO:0044699	single-organism process	1090	11.69	$< 1 \times 10^{-30}$
	GO:0044763	single-organism cellular process	988	10.61	$< 1 \times 10^{-30}$
	GO:0008152	metabolic process	912	9.61	$< 1 \times 10^{-30}$
	GO:0071704	organic substance metabolic process	885	9.28	$< 1 \times 10^{-30}$
	GO:0044238	primary metabolic process	849	8.86	$< 1 \times 10^{-30}$
	GO:0065007	biological regulation	912	9.88	$< 1 \times 10^{-30}$
	GO:0044237	cellular metabolic process	835	8.89	$< 1 \times 10^{-30}$
	GO:0050789	regulation of biological process	856	9.30	$< 1 \times 10^{-30}$
elim	GO:0045944	positive regulation of transcription fro...	77	0.98	$< 1 \times 10^{-30}$
	GO:0000122	negative regulation of transcription fro...	62	0.68	$< 1 \times 10^{-30}$
	GO:0006355	regulation of transcription, DNA-templat...	275	2.95	$< 1 \times 10^{-30}$
	GO:0007275	multicellular organism development	403	4.53	$< 1 \times 10^{-30}$
	GO:0008150	biological_process	1365	14.64	$< 1 \times 10^{-30}$
	GO:0055114	oxidation-reduction process	80	0.89	$< 1 \times 10^{-30}$
	GO:0008285	negative regulation of cell proliferatio...	59	0.59	$< 1 \times 10^{-30}$
	GO:0007165	signal transduction	441	5.01	$< 1 \times 10^{-30}$
	GO:0043547	positive regulation of GTPase activity	51	0.60	$< 1 \times 10^{-30}$
	GO:0030154	cell differentiation	295	3.35	$< 1 \times 10^{-30}$
weight	GO:0010467	gene expression	434	4.53	2.6×10^{-27}
	GO:0034645	cellular macromolecule biosynthetic proc...	406	4.23	7.7×10^{-24}
	GO:0006334	nucleosome assembly	36	0.11	9.3×10^{-19}
	GO:0032446	protein modification by small protein co...	97	0.84	3.2×10^{-17}
	GO:0044707	single-multicellular organism process	496	5.50	5.1×10^{-17}
	GO:0051291	protein heterooligomerization	30	0.10	7.4×10^{-17}
	GO:0007154	cell communication	486	5.46	1.3×10^{-16}
	GO:0022610	biological adhesion	169	1.58	6.0×10^{-15}
	GO:0019538	protein metabolic process	467	4.89	9.1×10^{-15}
	GO:0032200	telomere organization	25	0.13	7.3×10^{-14}
parent child	GO:0006334	nucleosome assembly	36	0.11	7.1×10^{-11}
	GO:0031497	chromatin assembly	38	0.12	9.8×10^{-11}
	GO:0006342	chromatin silencing	30	0.09	2.8×10^{-10}
	GO:0034728	nucleosome organization	39	0.13	1.3×10^{-9}
	GO:0040029	regulation of gene expression, epigeneti...	46	0.22	4.6×10^{-8}
	GO:0045814	negative regulation of gene expression, ...	30	0.10	8.8×10^{-8}
	GO:0016458	gene silencing	45	0.22	2.6×10^{-7}
	GO:0006333	chromatin assembly or disassembly	40	0.15	7.6×10^{-7}
	GO:0065004	protein-DNA complex assembly	41	0.19	8.2×10^{-7}
	GO:0071103	DNA conformation change	53	0.24	2.0×10^{-6}

Table S4: Top 10 biological process GO terms for genes located in hg19 loss hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in methods)

Algorithm	GO.ID	Term	Significant	Expected	p-value
classic	GO:0008150	biological_process	144	0.96	7.4×10^{-10}
	GO:0010243	response to organonitrogen compound	11	0.03	1.8×10^{-4}
	GO:0032874	positive regulation of stress-activated ...	5	0.01	5.1×10^{-4}
	GO:0070304	positive regulation of stress-activated ...	5	0.01	5.2×10^{-4}
	GO:1901698	response to nitrogen compound	11	0.04	5.6×10^{-4}
	GO:0006342	chromatin silencing	4	0.00	6.9×10^{-4}
	GO:0044699	single-organism process	78	0.61	7.6×10^{-4}
	GO:0045814	negative regulation of gene expression, ...	4	0.00	7.9×10^{-4}
	GO:1901700	response to oxygen-containing compound	15	0.07	8.8×10^{-4}
	GO:0009719	response to endogenous stimulus	15	0.07	9.2×10^{-4}
elim	GO:0008150	biological_process	144	0.96	$< 1 \times 10^{-30}$
	GO:0000122	negative regulation of transcription fro...	8	0.04	8.2×10^{-17}
	GO:0007186	G-protein coupled receptor signaling pat...	9	0.05	5.7×10^{-13}
	GO:0045944	positive regulation of transcription fro...	6	0.05	3.0×10^{-11}
	GO:0071230	cellular response to amino acid stimulus	4	0.01	8.3×10^{-11}
	GO:0055114	oxidation-reduction process	7	0.04	4.3×10^{-10}
	GO:0007608	sensory perception of smell	4	0.01	4.6×10^{-10}
	GO:0006355	regulation of transcription, DNA-templat...	22	0.15	4.0×10^{-9}
	GO:0046330	positive regulation of JNK cascade	4	0.01	4.0×10^{-9}
	GO:0007275	multicellular organism development	31	0.25	9.6×10^{-9}
weight	GO:0008150	biological_process	144	0.96	2.8×10^{-10}
	GO:0032874	positive regulation of stress-activated ...	5	0.01	5.1×10^{-4}
	GO:0006342	chromatin silencing	4	0.00	6.9×10^{-4}
	GO:0097094	craniofacial suture morphogenesis	2	0.00	1.7×10^{-3}
	GO:0050718	positive regulation of interleukin-1 bet...	2	0.00	2.6×10^{-3}
	GO:0006699	bile acid biosynthetic process	2	0.00	3.0×10^{-3}
	GO:0050810	regulation of steroid biosynthetic proce...	3	0.00	3.1×10^{-3}
	GO:0071230	cellular response to amino acid stimulus	4	0.01	3.1×10^{-3}
	GO:0030204	chondroitin sulfate metabolic process	2	0.00	5.0×10^{-3}
	GO:0001958	endochondral ossification	2	0.00	7.5×10^{-3}
parent child	GO:0001934	positive regulation of protein phosphory...	8	0.04	1.7×10^{-3}
	GO:0098927	vesicle-mediated transport between endos...	2	0.00	3.1×10^{-3}
	GO:0006342	chromatin silencing	4	0.00	3.1×10^{-3}
	GO:0070304	positive regulation of stress-activated ...	5	0.01	3.3×10^{-3}
	GO:0097094	craniofacial suture morphogenesis	2	0.00	3.4×10^{-3}
	GO:0070302	regulation of stress-activated protein k...	5	0.01	3.9×10^{-3}
	GO:0006346	methylation-dependent chromatin silencin...	2	0.00	3.9×10^{-3}
	GO:0045814	negative regulation of gene expression, ...	4	0.00	4.3×10^{-3}
	GO:0006066	alcohol metabolic process	5	0.01	4.6×10^{-3}
	GO:1901698	response to nitrogen compound	11	0.04	4.9×10^{-3}

Table S5: Top 10 biological process GO terms for genes located in mm10 gain hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in methods)

Algorithm	GO.ID	Term	Significant	Expected	p-value
classic	GO:0030154	cell differentiation	214	1.15	1.1×10^{-6}
	GO:0048869	cellular developmental process	229	1.24	1.3×10^{-6}
	GO:0051216	cartilage development	23	0.06	2.1×10^{-6}
	GO:0009888	tissue development	113	0.54	2.1×10^{-6}
	GO:0044767	single-organism developmental process	298	1.70	2.3×10^{-6}
	GO:0032502	developmental process	299	1.72	3.8×10^{-6}
	GO:0007275	multicellular organism development	258	1.46	4.8×10^{-6}
	GO:0048856	anatomical structure development	280	1.61	7.3×10^{-6}
	GO:0040007	growth	72	0.32	1.1×10^{-5}
	GO:0061448	connective tissue development	26	0.08	1.3×10^{-5}
elim	GO:0008150	biological_process	844	5.61	$< 1 \times 10^{-30}$
	GO:0045944	positive regulation of transcription fro...	64	0.32	$< 1 \times 10^{-30}$
	GO:0000122	negative regulation of transcription fro...	44	0.22	$< 1 \times 10^{-30}$
	GO:0006355	regulation of transcription, DNA-templat...	152	0.90	$< 1 \times 10^{-30}$
	GO:0008285	negative regulation of cell proliferatio...	41	0.19	$< 1 \times 10^{-30}$
	GO:0055114	oxidation-reduction process	41	0.26	$< 1 \times 10^{-30}$
	GO:0007275	multicellular organism development	258	1.46	$< 1 \times 10^{-30}$
	GO:0043066	negative regulation of apoptotic process	47	0.26	$< 1 \times 10^{-30}$
	GO:0001701	in utero embryonic development	32	0.13	$< 1 \times 10^{-30}$
	GO:0045893	positive regulation of transcription, DN...	84	0.41	$< 1 \times 10^{-30}$
weight	GO:1900271	regulation of long-term synaptic potenti...	7	0.01	3.7×10^{-5}
	GO:0045893	positive regulation of transcription, DN...	84	0.41	1.3×10^{-4}
	GO:0051216	cartilage development	23	0.06	1.8×10^{-4}
	GO:0045165	cell fate commitment	26	0.08	2.2×10^{-4}
	GO:0061180	mammary gland epithelium development	11	0.02	2.3×10^{-4}
	GO:0019827	stem cell population maintenance	17	0.05	2.7×10^{-4}
	GO:0003229	ventricular cardiac muscle tissue develo...	9	0.02	5.5×10^{-4}
	GO:0001704	formation of primary germ layer	12	0.03	5.6×10^{-4}
	GO:0031998	regulation of fatty acid beta-oxidation	5	0.01	6.0×10^{-4}
	GO:0007498	mesoderm development	13	0.03	6.2×10^{-4}
parent child	GO:0048869	cellular developmental process	229	1.24	7.5×10^{-5}
	GO:0051216	cartilage development	23	0.06	1.3×10^{-4}
	GO:1900373	positive regulation of purine nucleotide...	8	0.03	2.1×10^{-4}
	GO:0040007	growth	72	0.32	2.5×10^{-4}
	GO:0010463	mesenchymal cell proliferation	11	0.02	2.7×10^{-4}
	GO:0044767	single-organism developmental process	298	1.70	3.8×10^{-4}
	GO:0002281	macrophage activation involved in immune...	4	0.00	4.4×10^{-4}
	GO:1900271	regulation of long-term synaptic potenti...	7	0.01	7.8×10^{-4}
	GO:0048565	digestive tract development	16	0.04	1.1×10^{-3}
	GO:0032502	developmental process	299	1.72	1.2×10^{-3}

Table S6: Top 10 biological process GO terms for genes located in mm10 loss hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in methods)

Appendix C

Supplementary for Chapter 4

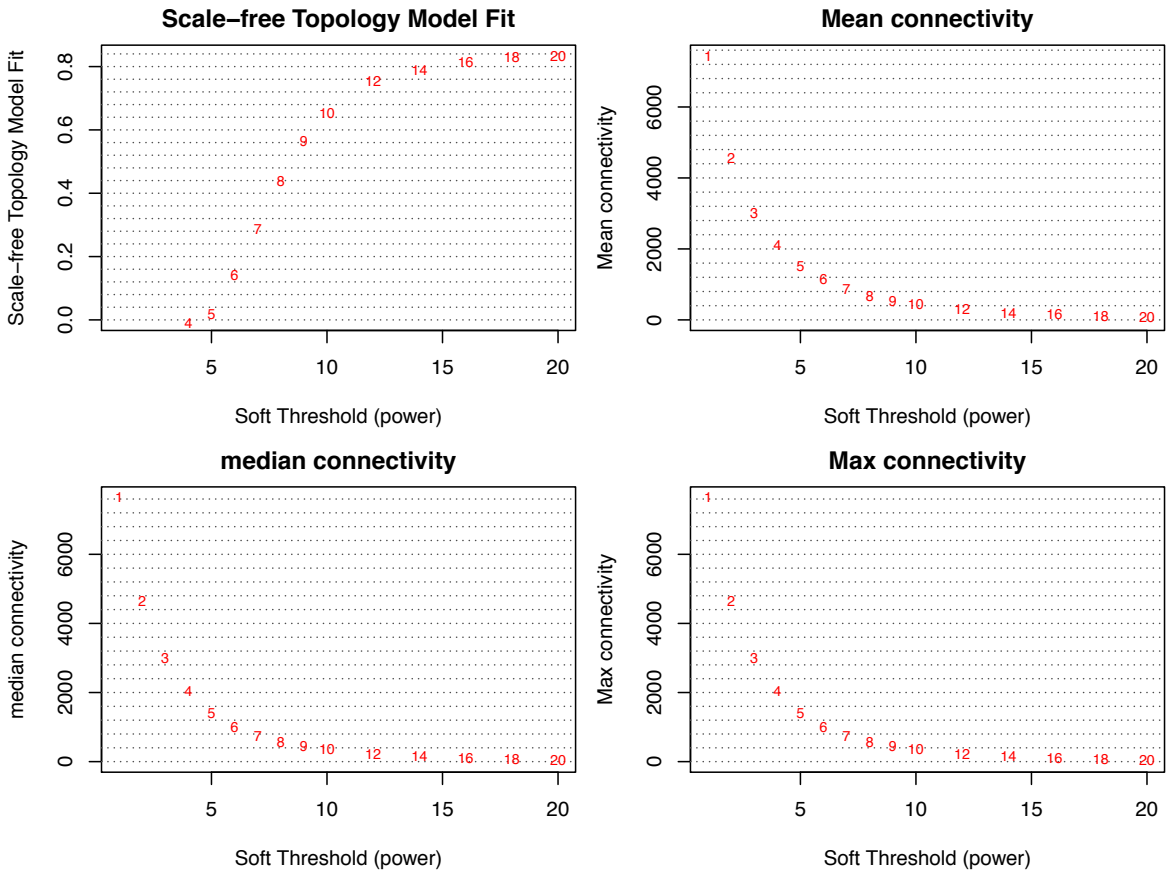


Figure S1: Statistics used to identify an appropriate soft threshold for our WGCNA anlysis.

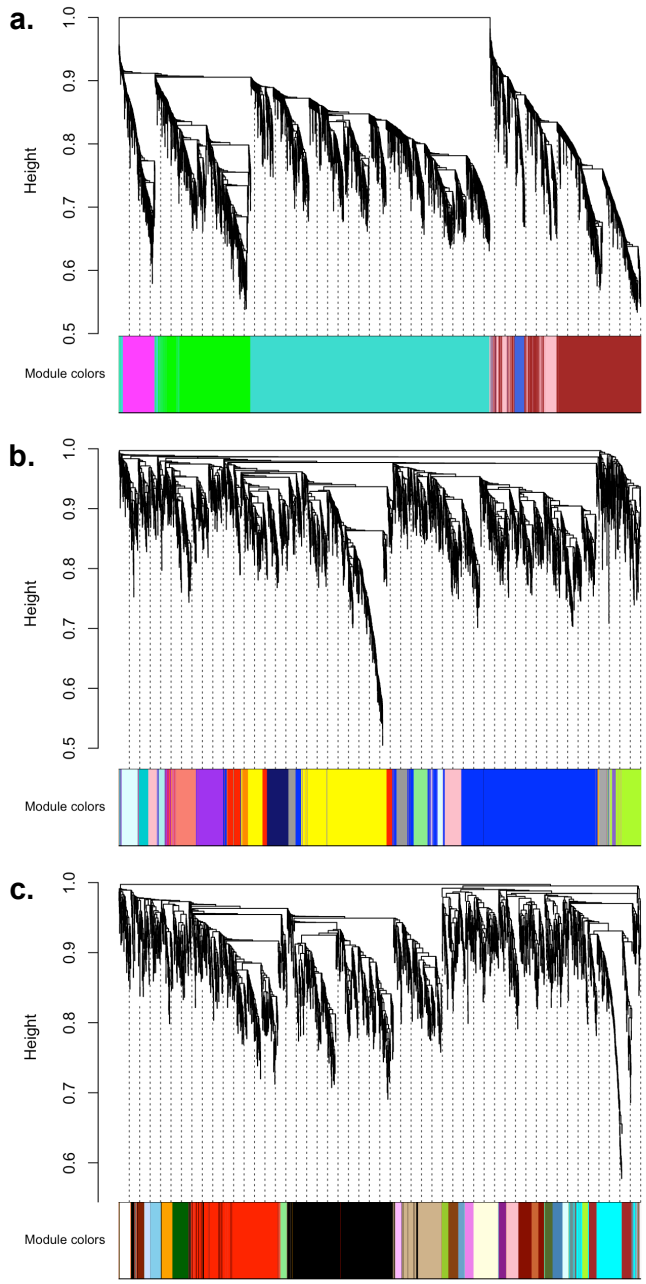


Figure S2: Shows the clustering pattern for all our samples along with information regarding study design and sequencing results. We find that variation across samples is mainly biological.

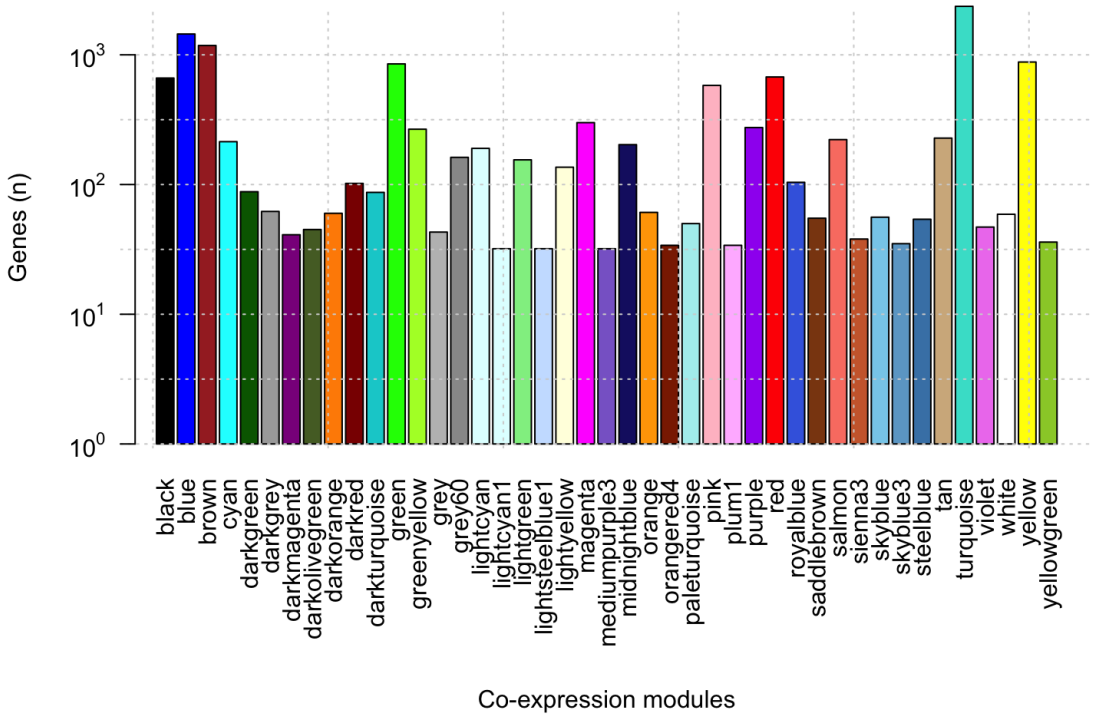


Figure S3: Hierarchical clustering dendrogram based on expression patterns of individual genes. Colours along the base of each dendrogram represent module membership. Modules were detected in a block-wise manner, where **a** is block 1, **b** is block 2 and **c** is block 3.

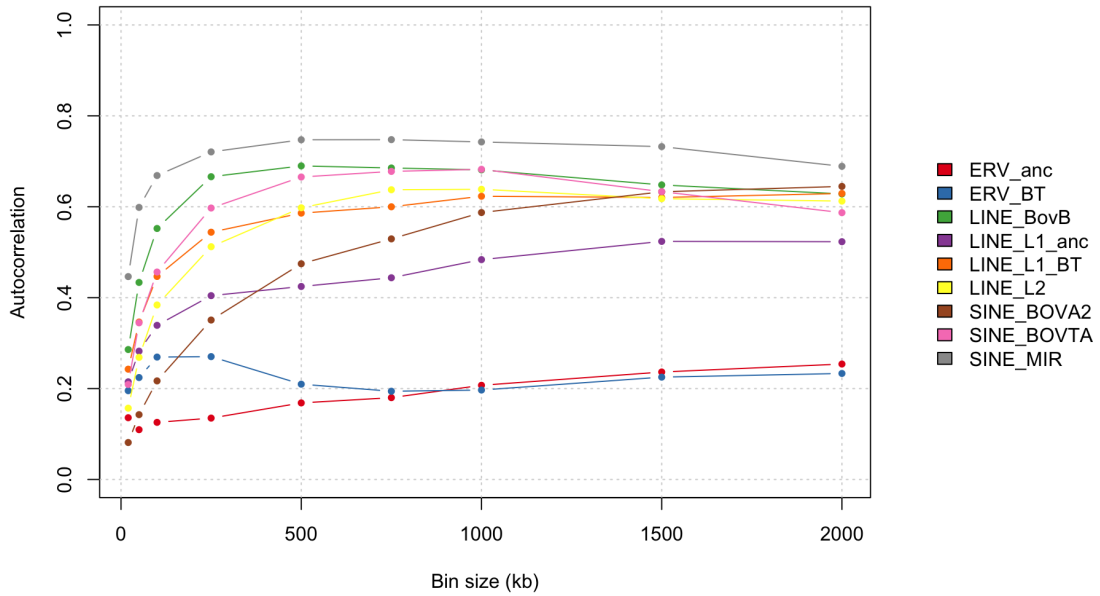


Figure S4: The number of genes that belong to each of our identified co-expression modules.

TE group	regex	Instances (n)	Annotated families (n)	Genome coverage (Mb)
ERV_anc	`ERV.*[^BT].$`	87280	202	15.84
ERV_BT	`ERV.*BT`	48356	19	14.91
LINE_BovB	`BovB`	506319	14	361.89
LINE_L1_anc	`L1.*[^BT]$`	713912	171	218.61
LINE_L1_BT	`^L1.*BT`	160317	4	108.20
LINE_L2	`^L2`	222964	24	42.31
SINE_BOVA2	`BOVA2`	209735	5	48.80
SINE_BOVTA	`Bov[^B].*`	415424	12	81.33
SINE_MIR	`MIR`	278673	6	36.29
Total		**2642980**	**457**	**928.18**

Table S1: TE groups extracted from bovine Censor output. Instances of individual TE that belong to each group were extracted using grep, which searched for a pattern match between one of our regular expressions and TE family names. The number of instances refers to the number of times a particular TE is found in the censor output rather than the number of individual insertions. This is because some TE insertions become fragmented over time and can be difficult to stitch together.

Appendix D

Supplementary for Chapter 5

Supporting Information

Chen et al. 10.1073/pnas.1519374113

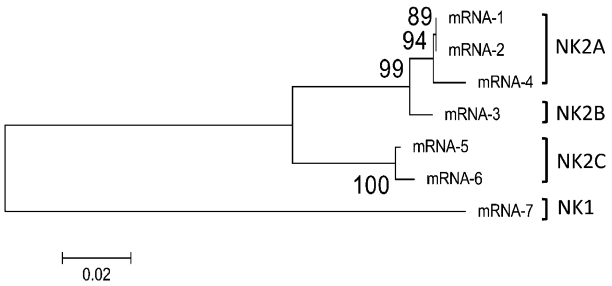
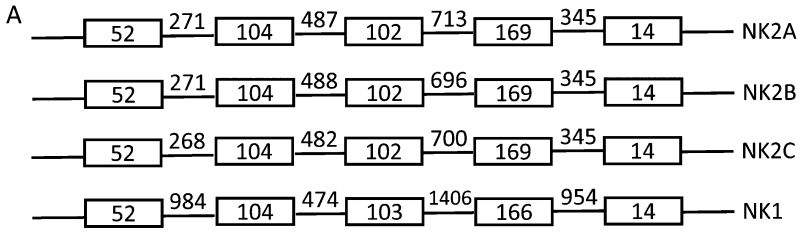


Fig. S1. Phylogenetic analysis of seven different bovine *NK-lysin*–related mRNA sequences (mRNA-1–7) from the NCBI nucleotide database. Four clades were formed and annotated as *NK1* and *NK2A*, *2B*, and *2C*. Bootstrap values are shown at branch points.



B

```
NK2A   MTSWAVLLIASVLLVAPGLAFSGLTPESHDQATAHLCDGDELCQGLDPEDPQGDLLLQGE 60
NK2B   -------I-------------------------T-Q------------------------ 60
NK2C   ---------T------------------------------------AL------------ 60
NK1    --------VT-----------------C---E----W---K-L---A--------V--R- 57

NK2A   ELGLLCGPCWMIMQHLMDKLGDQPDENTVIEVASKMCSKMRLLKGLCKSITKRFLRRIAA 120
NK2B   -----C-SCQR-I-----------------A---VCG--GP----C------------- 120
NK2C   --S-RC-SCRR-I-----------------E---VC---------C---M-K---T--E 120
NK1    --NQFC--CQK---M-KYMV-NR-SK-VI-H-T--VC---G-WSI-CNQMM-KY-N--SK 117
            *  *                            *         *
NK2A   DITAGKTSRVICEDIKMCKSKPVGFI 146
NK2B   ----------VC-----C-------- 146
NK2C   --V-----Q--CV-----C-------- 146
NK1    --M-R--PQA-CV---LC-R--A-L- 142
                  *     *
```

Fig. S2. Genomic structure and predicted amino acid sequence were compared among four bovine *NK-lysin* genes. (*A*) Size comparison of five exons and four introns. (*B*) Comparison of the predicted amino acid compositions. The amino acid sequence of *NK2A* was used as the reference. Six conserved cysteine residues are indicated.

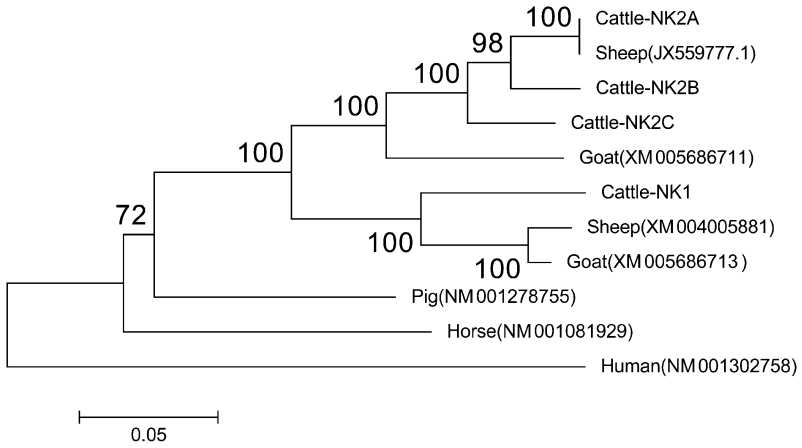


Fig. S3. Phylogenetic analysis of the full coding sequences of four bovine *NK-lysins* and *NK-lysin* orthologs in humans, pig, horse, sheep, and goat. The accession number for each sequence in the NCBI nucleotide database is indicated, and the bootstrap values are shown at branch points.

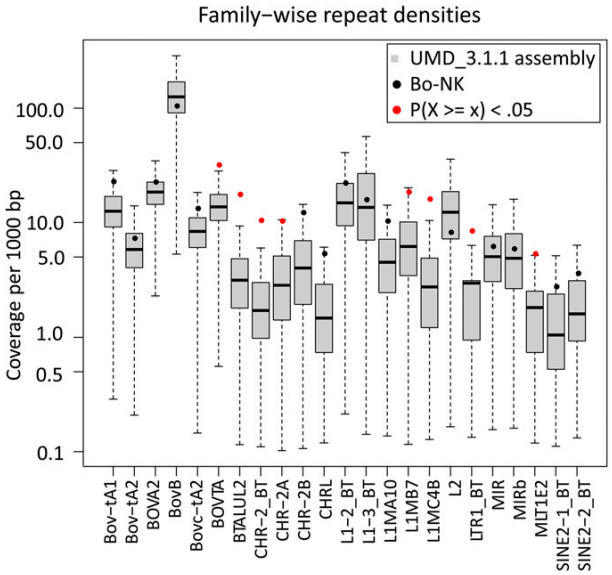


Fig. S4. Comparison of the repeat densities between the whole-genome assembly (UMD_3.1.1) and the assembled Bo-NK supercontig.

Table S1. *NK-lysin*–related sequences from the NCBI bovine nucleotide database

Sequence	Accession no.	Cluster
mRNA-1	XM_005192449	NK2A
mRNA-2	XM_005192450	NK2A
mRNA-3	BC114176	NK2B
mRNA-4	AY245798	NK2A
mRNA-5	BC114178	NK2C
mRNA-6	AY245799	NK2C
mRNA-7	NM_001046578	NK1

Table S2. Number of sequenced clones and different sequences obtained from each individual in the analysis of homozygous cattle

Animal ID	No. sequenced clones	No. different clones
2527	18	4
2796	30	3
2822	39	2
3850	30	2

Table S3. Summary of repeat 19274_elements within the Bo-NK supercontig

Class	Superfamily	Family	Frequency
LINEs			147
	L1		102
		HAL1	3
		L1_BT	3
		L1-2_BT	22
		L1-3_BT	13
		L1-BT	1
		L1MA10	12
		L1MB6_5	4
		L1MB7	13
		L1MC3	4
		L1MC4B	15
		L1MC5	1
		L1ME3C_3end	1
		L1ME3D_3end	3
		L1ME3E_3end	1
		L1ME5	1
		L1P_MA2	5
	L2		9
		L2	8
		L2B	1
	RTE		36
		BovB	36
SINEs			225
	SINE		24
		BCS	1
		BOVA2	23
	SINE2/tRNA		189
		Bov-tA1	28
		Bov-tA2	10
		Bov-tA3	4
		Bovc-tA2	11
		BOVTA	40
		BTALUL1	1
		CHR-2_BT	13
		CHR-2A	9
		CHR-2B	15
		CHRL	9
		CHRL1_BT	1
		MIR	11
		MIR3	3
		MIRb	13
		MIRc	5
		SINE2-1_BT	6
		SINE2-2_BT	8
		SINE2-3_BT	1
		THER1	1
	RTE		12
		BTALUL2	12
ERV			44
	ERV1		15
		BtERVF2_I	1
		ERV1-2-I_BT	2
		LTR1_BT	6
		LTR11_BT	3
		LTR39B_BT	1
		LTR39D_BT	1
		MER41_BT	1
	ERV2		1
		ERV2-1-LTR_BT	1
	ERV3		28
		LTR33C	1
		LTR67B	5

Table S3. Cont.

Class	Superfamily	Family	Frequency
		MLT1E2	9
		MLT1F	1
		MLT1F1	2
		MLT1J	2
		MLT1J1	1
		MLT1J2	3
		MLT1M	4
LTR			3
	Gypsy		3
		LTR88b	3
DNA transposon			8
	DNA transposon		1
		X25_DNA	1
	hAT		3
		Charlie13a	1
		MER91B	1
		UCON52	1
	Mariner/Tc1		4
		MER47B	2
		TIGGER5_B	1
		TIGGER5A	1

Totals in each category are underlined in the Frequency column. ERV, endogenous retrovirus; hAT, histone acetyltransferase; RTE, recombinational telomere elongation.

Table S4. Sequences and properties of four synthetic bovine NK-lysin peptides

Peptide	Sequence	Length, aa	Charge	Net charge, pH 7	Hydrophobicity, pH 6.8
NK1	VIIHVTSKVCSKMGLWSILCNQMMKKYLNR	30	+6	4.93	37.1
NK2A	TVIEVASKMCSKMRLLKGLCKSITKRFLRR	30	+8	7.82	31.43
NK2B	TVIEAASKVCGKMGPLKGLCKSITKRFLRR	30	+7	6.82	26.1
NK2C	TVIEEASKVCSKMRLLKGLCKSIMKKFLRT	30	+6	5.82	30.57

Table S5. Primer and probe information

Primer name	Forward, 5′→3′	Reverse, 5′→3′	Utilization
Bo-lysin	ACCCAGCACTCCCACTG	ACATACCTGGCTTGCTTTTG	Homozygotes analysis
JP-1	CTAAGTGGCCGGATTGTTGT	CAGGGTCTTCTCCTCTGACG	BAC assembly validation
JP-2	GAAATGCTCTCACAGCAACA	AATAGCAATGAAATGATGATGGT	BAC assembly validation
JP-3	AAAATGCTCTCACAGCAATGAA	AATAGCAATGAAATGATGGTAGCTG	BAC assembly validation
JP-4	GATAGTCTCCCCAACCAGTCAG	GAATTGCTGAGCTGGAAGAAGT	BAC assembly validation
BP-1	GCCTGCCTTCATGGAGTTTA	TGGCACAGGTAATGGGATAA	BAC assembly validation
Ex-NK1	CCAGCAAGAATGTCATCATCC	GTCCTTAGAGATGCGATTGAGATAC	Gene expression assay
Ex-NK2A	AGGAGAAGAGCTGGGCCTAC	GCTGATCTCCCAACTTGTCC	Gene expression assay
Ex-NK2B	GAGAATACCGTCATCGAGGC	TTGCACAGACCTTTCAGCG	Gene expression assay
Ex-NK2C	AATTTCTCCGTACCATCGCT	ATGAAACCTACTGGCTTGCTT	Gene expression assay
NK1-probe	CTTTGCAACCAGATGA		Gene expression assay
NK2A-probe	TCCTTGTTGGATGATAATG		Gene expression assay
NK2B-probe	TCCAAGGTGTGCGGC		Gene expression assay
NK2C-probe	AGGACATCGTAGCTGG		Gene expression assay
Gs-NK2B	CTGTTCATGCTGTTTCTTCCAT	TTGCACAGACCTTTCAGCG	*NK2B* deletion test